Neural Representation and Prior for Controllable 3D Digital Humans

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Garvita Tiwari

aus Varanasi/ Indien

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	31.10.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter/-in:	Prof. Dr. Gerard Pons-Moll
2. Berichterstatter/-in:	Prof. Dr. Peter Gehler
3. Berichterstatter/-in:	Assistant Prof. Georgios Pavlakos

"That the powerful play goes on, and you may contribute a verse." -Walt Whitman

Acknowledgements

Looking back on the past few years, it's incredible to see how far this journey has taken me. It has been filled with growth, learning, and unexpected challenges. Though the path was not without its trials, I never walked it alone. I want to take this moment to offer my heartfelt thanks to those whose presence and support helped shape this experience.

First and foremost, I would like to express my heartfelt gratitude to Prof. Dr. Gerard Pons-Moll for giving me the opportunity to work in this exciting area of research and for his guidance. I have learned so much from him, and his mentorship has shaped my scientific thinking in profound ways.

Next, I would like to thank the entire team at MPI-Informatics (D2), led by Prof. Dr. Bernt Schiele, for fostering such a friendly and inspiring environment for research. A special thanks to the outstanding administrative and IT support at MPI and the University of Tübingen, Connie and Jessi in particular, for making my life a million times easier with their swift and skillful navigation of German bureaucracy.

I have been incredibly lucky to find great collaborators throughout my journey. Thanks to Tony Tung for always being supportive of my research direction and for providing all the necessary support. I am immensely grateful to Nikolaos Sarafianos, perhaps the best collaborator one could ask for, always stepping in at the right time and taking care of the details that truly matter. Thanks also to Jan-Eric Lensen for broadening my research horizons. Special thanks to Prof. Dr. Peter Gehler and Prof. Dr. Georgios Pavlakos for reviewing my thesis, as well as Prof. Dr. Michael Black and Prof. Dr. Justus Thies for serving on my thesis committee.

I was fortunate to work with several new students in the lab and to learn from their energy. Working with Yannan and Dimitrije has been one of the most positive and enriching experiences I've had. I'm also thankful to Batuhan, Samuel, Daniel, Pankaj, and Margaret for the opportunity to supervise them; it has been a fulfilling experience in itself.

I would like to thank every member of the RVH group; I feel incredibly humbled to have worked alongside all of you. Thanks to Thiemo for his support and guidance. To Anna, Bharat, and Aymen – thank you for making the Covid days a bit more bearable with your warmth and presence. Special thanks to Riccardo, Julian, Ilya, Berna, Nadia, Nikita, Yuxuan – some of the brightest and kindest people I've had the privilege of knowing closely. Your brilliance, support, and kindness have shaped this journey in more ways than you can imagine. This acknowledgment wouldn't be complete without a shout-out to Freistil – the sacred temple of procrastination, where our research debates and beer-fueled therapy sessions went hand in hand. Cheers to science, friendship, and questionable life advice over pints!

I'm also grateful to Yao, Yuliang, and Qianli – your papers and work have been a consistent source of motivation throughout this journey.

I am deeply grateful for the friendships and connections I've built on this journey. To Divyanshu, Vikram, Chandrika, Shubho, Param, Atul, Aditi, Yash, Diala, Lorena, and Tarun, thank you for being the most incredible group of friends. Big shoutout to Divyanshu and Chandrika for basically adopting me – thanks for the second home, the actual best food in Germany, epic dinner parties, and movie nights. And to Arya – thank you for always keeping me grounded.

To my forever support system – Sonakshi, Shruti, Komal, Medhavi, Shreya, and Vedika, your achievements have been a constant source of inspiration, and your presence has filled my days with joy.

To my sister Harshita – thank you for always leading the way and testing the waters before me. To my parents – my foundation and my wings – thank you for your endless support and belief in me. And finally, thank you to Prakhar for your unwavering support, for being a patient listener through all my stories and complaints, and for introducing me to the world of computer vision.

Like all other thesis acknowledgments, this one, too, has been written a hundred times in my head and finally put on paper in a rush. It's impossible to name every person who has shaped this journey, but know that every gesture, conversation, and moment has meant a great deal. I feel incredibly fortunate and humbled to have shared this journey with all of you.

Abstract

Modeling realistic and controllable 3D digital humans is a central focus in computer vision and graphics, with broad applications in AR/VR, virtual try-on, entertainment, and embodied AI. Achieving high-fidelity avatars requires accurate modeling of how people look, dress, and move. However, this remains challenging due to high variability in clothing types, body shapes, poses, and their interdependencies. This dissertation addresses two core aspects of human digitization – how we dress and how we move – through novel representations, datasets, and learning-based methods that advance the state of 3D human modeling.

Our contributions to modeling human clothing address multi-faceted challenges across geometry representations, learning formulations, and practical applications. We explore how clothing and body geometry are influenced by size, shape, pose, and style, proposing specialized representations and learning frameworks—supported by high-quality datasets—to enable more robust and generalizable models of clothed humans.

We introduce SIZER, the first model to capture clothing deformations based on clothing size and body shape using a mesh-based representation. SizerNet models the subtle, non-linear effects of size and shape on clothing fit and drape – factors often overlooked in prior work. To support this, we curate a large-scale, high-quality dataset of 3D scans with diverse subjects, clothing types, and multiple sizes per item. Rich annotations, including clothing categories, segmentation, registrations, and accurate body shapes, enable realistic, data-driven modeling. Additionally, we introduce ParserNet, the first learning-based method to recover detailed clothing meshes and body shape from a single input mesh. Unlike prior methods, it allows editing clothing or body shape without requiring segmentation or registration at test time. Together, the dataset and models form a strong foundation for future work in clothed human reconstruction, generation, and size-sensitive modeling.

While mesh-based approaches excel in template-specific garments, they struggle with non-rigid deformations, topological diversity, and require costly registration and processing pipelines. To address this, we propose Neural-GIF, an animatable model of clothed humans learned directly from raw scans using neural implicit representations. Inspired by template-based decomposition of motion into articulation and non-rigid deformation, we generalize this idea to implicit shape learning. We learn a pose-dependent backward skinning field to map input points to a canonical space, followed by a deformation field in the canonical space to capture non-rigid, pose-dependent deformation. This enables modeling of highly complex clothing and soft-tissue dynamics. Neural-GIF achieves fine-grained details, shape-aware deformations, and outperforms prior methods in flexibility and accuracy. We also introduce a 4D scan dataset with loose, complex clothing to support further research in this area.

Prior work often models clothing and body as a single surface or focuses on isolated clothing styles, failing to capture the diversity and layered nature of real-world clothing. To address this, we propose TailorDiff, a unified, controllable, diffusion-based point cloud

model that serves as an SMPL-like foundation model for clothing. By leveraging a point cloud representation, TailorDiff is topology-agnostic, enabling unified modeling across all categories of clothing. A hierarchical, body-aware latent space supports generation and editing conditioned on style, body shape, and fine-grained attributes like sleeve or hem length. To train this large-scale model, we curate a rich and diverse dataset by aggregating synthetic clothing scans from multiple sources. TailorDiff also readily supports inverse problems, such as fitting layered clothing to partial scans – pushing the boundaries of SMPL-based human modeling towards more expressive, realistic representations of humans in layered clothing.

Beyond modeling how we dress, this dissertation also addresses how we move. Pose priors are crucial for generating realistic new poses and for reconstructing accurate poses from noisy or partial observations. Existing priors rely on Gaussian assumptions, which bias estimates towards the mean and distort the pose manifold. We introduce PoseNDF, a novel formulation for learning pose priors using a Neural Distance Field, without relying on explicit density modeling or prior assumptions. PoseNDF extends implicit surface modeling in 3D to high-dimensional pose space, learning the valid pose manifold from sets of plausible and implausible poses via a neural network. The zero level set of the learned manifold defines valid poses, enabling projection using the distance and its analytic gradient. Using PoseNDF as prior yields state-of-the-art results in motion denoising, pose completion, and single-image 3-D pose estimation while increasing sample diversity.

This thesis advances the field of 3D digital humans by pushing the frontiers of human-clothing models and human pose understanding. We provide an in-depth analysis of our ideas and position our work in the context of related research, including approaches that emerged concurrently. We conclude with an outline of future directions, aiming not only to address current limitations but also to further advance the field. As the demand for immersive digital environments continues to grow, we envision our contributions serving as a foundation for building expressive, generalizable, and fully interactive 3D digital humans.

Zusammenfassung

Die Modellierung von realistischen und kontrollierbaren digitalen 3D Menschen ist in zentrales Thema des maschinellen Sehens und der Computerrgrafik mit vielfältigen Anwendungen in der AR/VR, virtuellen Anprobe, Unterhaltungsindustrie und verkörperter KI.Die Erstellung hochauflösender Avatare erfordert eine präzise Modellierung der Erscheinung, der Kleidung und der Bewegungen von Menschen. Dies ist nach wie vor herausfordernd aufgrund der hohen Variabilität von Kleidungstypen, Körperformen, Körperhaltungen und deren wechselseitigen Abhängigkeiten. Diese Dissertation behandelt zwei zentrale Aspekte der menschlichen Digitalisierung – wie wir uns kleiden und wie wir uns bewegen – durch neuartige Repräsentationen, Datensätze und lernbasierte Methoden, die den Stand der 3D Modellierung des Menschen vorantreiben.

Unsere Beiträge zur Modellierung menschlicher Bekleidung greifen vielfältige Herausforderungen in Bezug auf Geometrierepräsentationen, Lernformulierungen und praktische Anwendungen auf. Wir untersuchen, wie Kleidung und Körpergeometrie durch Größe, Form, Pose und Stil beeinflusst werden, und schlagen spezialisierte Repräsentationen und Lernframeworks vor – gestützt durch hochwertige Datensätze – um robustere und generalisierbare Modelle bekleideter Menschen zu ermöglichen.

Wir stellen SIZER vor – das erste Modell, das Kleidungsdeformationen basierend auf Kleidergröße und Körperform mit Hilfe einer mesh-basierten Repräsentation erfasst. SizerNet modelliert subtile, nichtlineare Effekte von Größe und Form auf Passform und Fall der Kleidung – Faktoren, die in früheren Arbeiten oft übersehen wurden. Dazu erstellen wir einen groß angelegten, hochwertigen Datensatz aus 3D-Scans von diversen Subjekten, welche verschiedene Kleidungstypen in verschiedenen Größen tragen. Reichhaltige Annotationen, darunter Kleidungskategorien, Segmentierungen, Registrierungen und präzise Körperformen, ermöglichen eine realistische, datengetriebene Modellierung. Außerdem präsentieren wir ParserNet, die erste lernbasierte Methode zur Wiederherstellung detaillierter Kleidungsmeshes und Körperformen aus einem einzelnem Eingabemesh. Im Gegensatz zu früheren Ansätzen erlaubt es das Editieren von Kleidung oder Körperform ohne Segmentierung oder Registrierung zur Testzeit. Zusammen bilden der Datensatz und die Modelle ein starkes Fundament für zukünftige Arbeiten zur Rekonstruktion, Generierung und größenabhängigen Modellierung bekleideter Menschen.

Während mesh-basierte Ansätze bei vorlagenspezifischer Kleidung gut funktionieren, stoßen sie bei nichtlinearen Verformungen, topologischer Vielfalt und aufwändigen Registrierungs- und Verarbeitungsprozessen an ihre Grenzen. Um dem zu begegnen, schlagen wir Neural-GIF vor – ein animierbares Modell bekleideter Menschen, das direkt aus unverarbeiteten Scans durch neuronale implizite Repräsentationen gelernt wird. Inspiriert von der vorlagenbasierten Zerlegung von Bewegung in Artikulation und nichtlineare Verformung, generalisieren wir diesen Ansatz auf implizites Lernen von Formen. Wir lernen ein posenabhängiges Rück-

transformationsfeld, das Eingabepunkte in einen kanonischen Raum abbildet, gefolgt von einem Deformationsfeld in diesem Raum zur Erfassung nichtlinearer, posenabhängiger Verformungen. Dies ermöglicht die Modellierung hochkomplexer Kleidung und Weichgewebedynamik. Neural-GIF kann feine Details und form abhängige Verformungen darstellen und übertrifft frühere Methoden in Flexibilität und Genauigkeit. Zudem stellen wir einen 4D-Scan-Datensatz mit lockerer, komplexer Kleidung bereit, der weitere Forschung in diesem Bereich unterstützt.

Frühere Arbeiten modellieren Kleidung und Körper oft als eine einzelne Oberfläche oder konzentrieren sich auf isolierte Kleidungsstile und erfassen so weder die Vielfalt noch die Mehrlagigkeit realer Bekleidung. Um dies zu beheben, schlagen wir TailorDiff vor – ein einheitliches, steuerbares, diffusionbasiertes Punktwolkenmodell, das als SMPL-ähnliches Grundmodell für Kleidung dient. Durch die Nutzung der Punktwolkenrepräsentation ist TailorDiff topologieunabhängig und erlaubt eine einheitliche Modellierung aller Kleidungskategorien. Ein hierarchischer, körpersensitiver latenter Raum ermöglicht die Generierung und Bearbeitung basierend auf Stil, Körperform und feineren Attributen wie Ärmellänge oder Saumlänge. Zum Trainieren dieses großen Modells erstellen wir einen umfangreichen und vielfältigen Datensatz indem wir synthethische Kleidungdsscans aus verschiedenen Quellen zusammentragen. TailorDiff unterstützt auch Inverse Probleme wie das Anpassen mehrlagiger Kleidung an partielle Scans – ein Schritt hin zu ausdrucksstärkeren, realistischeren Darstellungen von Menschen in mehrlagiger Kleidung.

Neben der Modellierung, wie wir uns kleiden, widmet sich diese Dissertation auch der Frage, wie wir uns bewegen. *Pose-priors* sind entscheidend für die realistische Generierung neuer Posen sowie die Rekonstruktion genauer Posen aus verrauschten oder unvollständigen Beobachtungen. Bestehende Priors basieren auf Gaußschen Annahmen, die Schätzungen zum Mittelwert hin verzerren und die Posenmannigfaltigkeit verzerren. Wir stellen PoseNDF vor, eine neuartige Formulierung zum Lernen von mittels *Neural Distance Fields*, ohne explizite Dichteannahmen oder vorherige Modellannahmen. PoseNDF erweitert implizite 3D-Oberflächenmodellierung auf hochdimensionale Posenräume, indem es gültige und ungültige Posenmengen über ein neuronales Netz lernt. Die Nullniveaufläche der gelernten Mannigfaltigkeit definiert gültige Posen und erlaubt Projektionen mittels Distanz und analytischem Gradienten. Die Verwendung von PoseNDF als Prior führt zu neuen Bestleistungen in Bewegungsentrauschung, Posenvervollständigung und 3D-Pose-Schätzung aus Einzelbildern – bei gleichzeitig höherer Diversität der erzeugten Posen.

Diese Dissertation erweitert das Feld der 3D-Digitalmenschen, indem sie die Grenzen von Kleidung- und Bewegungsmodellen verschiebt. Wir bieten eine tiefgehende Analyse unserer Ansätze und positionieren unsere Arbeit im Kontext verwandter und gleichzeitig entstandener Forschungsrichtungen. Abschließend skizzieren wir zukünftige Forschungsrichtungen, die aktuelle Einschränkungen adressieren und das Feld weiter vorantreiben sollen. Angesichts der wachsenden Nachfrage nach immersiven digitalen Umgebungen sehen wir unsere Beiträge als Grundlage für die Entwicklung ausdrucksstarker, generalisierbarer und vollständig interaktiver 3D-Digitalmenschen.

Table of contents

1 Introduction **1**
- 1.1 Motivation . . . 1
  - 1.1.1 Human in Clothing . . . 2
  - 1.1.2 Human Pose Priors . . . 3
- 1.2 Contributions . . . 4
  - 1.2.1 SIZER: Parsing and Learning Size Sensitive 3D Clothing . . . 4
  - 1.2.2 NeuralGIF for Animating People in Clothing . . . 5
  - 1.2.3 TailorDiff: Unified and Expressive Generative Model for 3D Clothing 6
  - 1.2.4 PoseNDF: Modeling Human Pose Manifolds with NDFs . . . 6
- 1.3 Publications and Works . . . 7
- 1.4 Outline . . . 8

2 Background **11**
- 2.1 Digital Avatars . . . 11
  - 2.1.1 Linear Blend Skinning . . . 11
  - 2.1.2 Body Models . . . 13
- 2.2 3D Representations for Human in Clothing . . . 15
  - 2.2.1 Meshes . . . 16
  - 2.2.2 Point Clouds . . . 16
  - 2.2.3 Implicit Representations . . . 17
- 2.3 Generative Models . . . 18
  - 2.3.1 Diffusion Models . . . 18
  - 2.3.2 Latent Diffusion Models . . . 19

3 Related Work **21**
- 3.1 Human in Clothing . . . 21
  - 3.1.1 Digital Avatars and Clothing: A Historical Overview . . . 21
  - 3.1.2 Human in Clothing Datasets . . . 23
  - 3.1.3 Mesh-Based Approaches . . . 27
  - 3.1.4 Neural Fields for 3D Human Modeling . . . 30
  - 3.1.5 Point-Based Representations of Humans in 3D . . . 32
  - 3.1.6 Layered and Disentangled Human in Clothing . . . 35
- 3.2 Human Pose Priors . . . 36
  - 3.2.1 Pose and Motion Priors . . . 36
  - 3.2.2 Neural Implicit for Learning Data Manifold . . . 37

I Human in Clothing **39**

4 SIZER: Parsing and Learning Size Sensitive 3D Clothing **41**
4.1 Introduction . . . 42
4.2 Mesh-based Clothed Human Representation . . . 45
4.2.1 Clothed Human Representation via Vertex Displacements . . . 46
4.2.2 Clothing Representation via SMPL+G Submeshes . . . 46
4.3 SIZER Dataset . . . 47
4.3.1 Dataset Overview . . . 47
4.3.2 Clothing Size Variation in Real World . . . 49
4.4 3D Human Parsing . . . 50
4.4.1 3D Clothing Parsing . . . 51
4.4.2 Body Shape under Clothing . . . 52
4.4.3 Training and Implementation Details . . . 53
4.5 Clothing Resizing Model . . . 55
4.5.1 Sizing deformation model . . . 55
4.5.2 Training and Implementation Details . . . 55
4.6 Experiments and Results . . . 56
4.6.1 3D Clothing Parsing and Shape under Clothing . . . 57
4.6.2 3D Clothing Resizing . . . 58
4.6.3 Draping on Unseen Body . . . 67
4.6.4 Ablation study . . . 67
4.7 Conclusion . . . 68
4.7.1 Limitations and Discussions . . . 70
4.7.2 Ethical Considerations . . . 70
4.7.3 Future Work and Impact . . . 71

5 NeuralGIF for Animating People in Clothing **73**
5.1 Introduction . . . 74
5.2 Generalized Implicit Functions(GIF) . . . 77
5.2.1 GIF for Rigid Transformations . . . 78
5.2.2 GIF for Displacement Field-Based Transformation . . . 78
5.3 Animating People in Clothing . . . 79
5.3.1 Canonical Mapping Network . . . 80
5.3.2 Displacement Field Network . . . 81
5.3.3 Canonical SDF and Normal Prediction . . . 81
5.3.4 Multi-Subject Model using Neural-GIF . . . 82
5.3.5 Training and Implementation Details . . . 82
5.4 Skinning & Displacement Fields: Bridging Classical and Implicit Models . . . 83
5.4.1 Forward Skinning and Backward Skinning . . . 84
5.4.2 Displacement Field and Blend Shapes . . . 86
5.5 ClothSeq Dataset . . . 88
5.6 Experiments and Results . . . 90
5.6.1 Reanimating People in Complex Clothing . . . 91

5.6.2 Modeling Soft-tissue Dynamics . . . 94
5.6.3 Comparisons and Ablation Studies . . . 95
5.6.4 Multi-Subject Model . . . 100
5.7 Conclusion . . . 101
5.7.1 Limitation and Discussions . . . 102
5.7.2 Future Work and Impact . . . 103

6 TailorDiff: Unified and Expressive Generative Model for 3D Clothing 105
6.1 Introduction . . . 106
6.2 A Unified and Expressive Generative Model for 3D Clothing . . . 109
6.2.1 Hierarchical Latent Space . . . 110
6.2.2 Latent Point Diffusion Model . . . 112
6.3 Dataset . . . 113
6.3.1 Synthetic Clothing Datasets . . . 113
6.3.2 Dataset Processing and Alignment . . . 114
6.4 Experiments and Results . . . 116
6.4.1 Generation using TailorDiff . . . 116
6.4.2 Cloth Editing using TailorDiff . . . 120
6.4.3 Body Shape Editing using TailorDiff . . . 126
6.4.4 Clothing Reconstruction . . . 129
6.4.5 Fitting TailorDiff to Scans . . . 133
6.4.6 Ablation . . . 135
6.5 Conclusion . . . 137
6.5.1 Limitation and Discussions . . . 138
6.5.2 Future Work . . . 138

II Human Pose Priors 141

7 PoseNDF: Modeling Human Pose Manifolds with NDFs 143
7.1 Introduction . . . 144
7.2 Method . . . 147
7.2.1 Quaternions as Representation of SO(3) . . . 147
7.2.2 Hierarchical Implicit Neural Function . . . 148
7.2.3 Loss functions . . . 149
7.2.4 Sampling via Projection on Manifold . . . 149
7.2.5 Data Preparation . . . 150
7.2.6 Training and Implementation details . . . 150
7.3 Learning and Sampling Data Manifold . . . 151
7.3.1 Energy Based Implicit Manifold . . . 151
7.3.2 Sampling in NDF and Diffusion . . . 151
7.4 Experiments and Results . . . 152
7.4.1 Experimental Setup . . . 153
7.4.2 Denoising Mocap Data . . . 153

7.4.3 Fitting to Partial Data . . . 157
7.4.4 3D Pose Estimation from Images . . . 159
7.4.5 Pose Generation . . . 162
7.4.6 Pose Interpolation . . . 163
7.4.7 Pose-NDF vs. Gaussian Assumption Models . . . 165
7.5 Conclusion . . . 168
7.5.1 Limitation and Discussions . . . 168
7.5.2 Future Work and Impact . . . 168

8 Conclusion **171**
8.1 Key Insights . . . 173
8.1.1 Human in Clothing . . . 173
8.1.2 Human Pose Priors . . . 176
8.2 Future Direction . . . 177
8.3 Social Implications . . . 181
8.4 A Broader View on Topic . . . 181

List of figures **183**

List of tables **191**

References **195**

1 Introduction

The ability to express our individuality through movement, attire, and interaction is quintessentially human. We reveal who we are not only through words, but through our behavior, the clothing and styling we choose, the subtleties of our gestures, and the ways we interact. Inherently, our world is shaped by these dynamic, deeply personal behaviors, which encompass a rich palette of human diversity.

Today, the same individuality extends into the digital realm. Avatars, social media personas, digital content creation, and immersive AR/VR experiences have become new canvases for self-expression, allowing us to carry our unique mannerisms, styles, and preferences into virtual spaces. Capturing this breadth of expression with fidelity, both in motion and in clothing, is crucial for creating digital worlds that feel authentic, inclusive, and genuinely human. As a result, research and studies on generating and controlling digital humans have tremendous applications, including telepresence, virtual reality, gaming, entertainment, fashion, and e-commerce. In this thesis, we present works that advance the field of digital humans by making significant contributions to human clothing and human pose understanding.

In Section 1.1, we present the motivation and key research questions addressed in the domain of human clothing models and human pose priors. A comprehensive overview of the publications and technical contributions resulting from this thesis is provided in Section 1.2. Finally, in Section 1.4 we provide an outline of the overall thesis structure.

1.1 Motivation

In response to the growing emphasis on individuality in digital environments, there is a pressing need for digital human models that are realistic, controllable, and application-ready. As digital presence becomes increasingly expressive, capturing human appearance and motion with high fidelity has become foundational to applications ranging from immersive storytelling and telepresence to fashion, VR, and e-commerce. In e-commerce alone, inaccuracies in clothing representation and pose result in significant economic losses due to returns, highlighting the importance of accurate and efficient clothing and human modeling. With the rise of AR/VR, digital humans are increasingly essential for immersive experiences, interactive storytelling, and AI-driven fashion design. However, accurately capturing, generating, and animating human bodies and clothing presents unique challenges due to the complexity and diversity of body shape variations, clothing topologies, and deformations caused by movement and interaction.

Realistic modeling of human clothing and accurate pose estimation are not merely aesthetic pursuits; they are fundamental to advancing critical real-world applications. Primarily, human modeling has been driven by parametric models, such as SMPL, which offer structured and interpretable representations. However, these models are limited in capturing detailed clothing and non-rigid deformations, often resulting in unrealistic poses due to the lack of constraints. In this thesis, we present work that pushes the boundaries of 3D human modeling and understanding beyond prior efforts. We make valuable contributions through the introduction of large-scale datasets and data-driven, learning-based methods for clothing modeling, ranging from capturing fine details to proposing a generalizable and unified foundational model of clothing. Additionally, we develop pose priors that constrain the realism of human poses.

Specifically, our contributions span two major directions: realistic *Human in Clothing* models and *Human Pose Priors*. For *Human in Clothing*, we explore various 3D representations to model diverse clothing behaviors, including size-sensitive deformations and highly complex dynamic deformations, culminating in a generalizable, unified foundational model of clothing, analogous to SMPL for human bodies. For *Human Pose Priors*, we introduce a novel formulation for learning the data manifold of plausible poses and draw connections to more general data prior learning frameworks.

1.1.1 Human in Clothing

Realistic digital human modeling requires not only accurate reconstruction of the human body [PCG*19, HLRB12, PMRMB15], but also detailed and physically plausible clothing. As previously highlighted, modeling humans and clothing together is critical for applications such as virtual try-ons, character animation, and immersive AR/VR experiences, where the realism of the interaction between the body and clothing plays a key role in enhancing visual fidelity and user engagement. To achieve such realism and physical plausibility, it is essential to capture subtle clothing details, including wrinkles, folds, and multi-layered structures, and to model how these elements respond to variations in body shape, clothing style, fit, and dynamic human poses. Recent advances in large-scale image and video generation models [KLL*24, MYCL24, WAW*25] have demonstrated remarkable capabilities in synthesizing realistic human avatars with complex clothing shapes and motion. These models benefit from powerful 2D/video priors and massive real-world datasets, but they often lack native 3D representations, expressive controllability, and physical realism.

Beyond realism, visual fidelity, and physical plausibility, modeling clothing via learning frameworks that can capture the vast diversity of real-world clothing styles and structures remains an open challenge. This includes not only modeling clothing behavior in response to underlying factors, but also topological variation and design-driven features. Developing models and representations that generalize across such diversity is crucial for building a scalable and versatile solution for 3D modeling.

These challenges highlight a broader need for more principled and expressive 3D modeling frameworks that can complement the generative power of 2D methods while enabling fine-grained control and maintaining consistent physical structure.

Building on these motivations, we highlight fundamental open questions in the research on clothed digital humans, spanning data acquisition, learning physically grounded clothing behavior, and identifying effective representations suitable for both learning and applications. In Part I of this thesis, we address these challenges through our contributions [TBTPM20, TSTPM21, TXL*25], which focus on learning-based 3D clothing models. Specifically, we address key questions that continue to shape the field: *Should we rely on explicit mesh-based models that offer structured and application compatible representation, despite their limitations in handling complex topologies and layered garments? Can neural implicit representations, with their flexibility and capacity to encode fine geometric details, serve as a better alternative, despite their challenges in simulation and rendering pipelines? Or do point cloud-based representations offer a compelling compromise, balancing geometric richness, efficiency, and learning compatibility?* Additionally, we explore the extent to which synthetic data can be leveraged to train robust and generalizable models, especially in scenarios where real data is scarce. Addressing these challenges is crucial for advancing 3D digital human modeling, paving the way for more realistic, robust, and controllable human avatars in diverse application scenarios. In the following sections, we present an overview of the contributions made throughout this thesis toward modeling human clothing.

1.1.2 Human Pose Priors

In Part II of this thesis, we discuss another crucial aspect of human modeling and the creation of digital avatars: the way we move, or more specifically, how we pose. Realistic and physically accurate human motion, pose capture, and generation play a central role in understanding human behavior and interactions within 3D environments. Applications such as immersive VR/AR, character animation, content creation, robotics, and human-computer interaction all depend on the ability to reconstruct or generate realistic human movement. Despite considerable progress, state-of-the-art capture techniques such as marker-based, IMU-based, or RGB/RGB-D reconstructions, often produce physically implausible or visually unrealistic human poses. These errors include artifacts such as foot skating, joint dislocations, self-intersections, local joint artifacts, and temporal jitter, especially in the presence of noise or occlusions. These artifacts degrade the realism and usability of human motion data in downstream applications. To mitigate these issues, pose priors are often used to regularize motion and pose, thereby ensuring the plausibility and realism of reconstructed human motion or pose.

This raises core research questions: How can we model human pose and motion in a way that captures the full diversity and physical plausibility of real-world data? Prior works have shown to constrain human poses using handcrafted priors [ENNE12] or data-driven priors with Gaussian Assumptions [BKL*16, PCG*19]. *Can we move beyond traditional Gaussian-based priors and define a continuous, data-driven manifold of valid poses that supports optimization and generalization?* Our work contributes to this trajectory by introducing PoseNDF, a neural distance field-based pose prior that defines plausible human poses as the zero-level set of an implicit function, enabling meaningful gradients, improved realism, and better diversity. This approach has not only shown strong results for pose optimization but

also inspired broader use of neural distance fields across motion [DFZ*24], vision, and more generic data [RLGCC23].

1.2 Contributions

This thesis presents significant contributions to the fields of human clothing and human pose through the development of novel datasets, methods, and formulations. These contributions address current limitations and offer comprehensive solutions for emerging tasks, such as 3D size-sensitive clothing modeling, as well as improved datasets and methodologies for established problems, including animating people in clothing and learning pose priors. Each contribution advances the state of the art by achieving a higher degree of realism, which was previously unattainable without the proposed formulations and contributions.

1.2.1 SIZER: Parsing and Learning Size Sensitive 3D Clothing

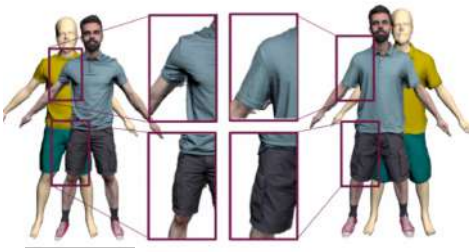


Fig. 1.1 **SIZER:** A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing.

In **Chapter 4**, we present "SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing" [TBTPM20]. In this work, we introduce the first comprehensive approach to size-sensitive clothing deformation in 3D, addressing the previously overlooked problem of how real-world clothings change with variations in clothing size and body shape. Unlike prior methods that either modeled clothing as a single-layer surface or used synthetic templates, SIZER enables explicit control over clothing fit and sizing, laying the foundation for realistic and personalized virtual try-on systems. The work extends beyond traditional shape- and pose-dependent deformation by focusing on subtle, real-world variations caused by clothing sizing, a critical factor for e-commerce, fashion analysis, and AR/VR applications.

We introduce a novel dataset–the first large-scale dataset specifically designed to capture real-world variations in clothing sizes. It includes approximately 2,000 high-quality 3D scans of 100 diverse subjects, each wearing casual clothing items in up to four sizes (S, M, L, XL). Each scan is accompanied by detailed annotations, including SMPL-based registrations, clothing part segmentation, clothing category labels, and size information. This dataset enables a rigorous study of clothing deformation as a function of clothing fit and body shape and serves as a valuable benchmark for 3D reconstruction [XXMPM24], segmentation [ATO*24, SDK*25], and layered-clothing modeling tasks [PLC*24, DLLG*23].

Built upon this dataset, our proposed method combines two novel neural architectures. ParserNet infers a multilayered representation, separating the body, upper, and lower garment from a single-layer mesh input (e.g., an SMPL-aligned scan). ParserNet was the first neural-network-based method capable of extracting semantically meaningful layers, replacing

the complex manual segmentation pipelines used in prior works [BTTPM19, PMPHB17a]. SizerNet, on the other hand, models clothing deformation conditioned on body shape and clothing size. Because clothing deformation as a function of size is highly non-linear and fine-grained, we propose an autoencoder-based approach to learn this deformation in canonical space. This combination of a novel dataset and a scalable, learning-based method enables controllable 3D clothing modeling and editing, significantly advancing the field of digitizing clothed humans.

1.2.2 NeuralGIF for Animating People in Clothing

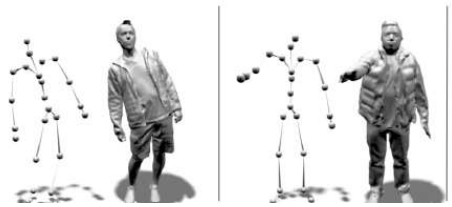


Fig. 1.2 **NeuralGIF:** Animating people in complex clothing.

In **Chapter 5**, we present "Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing" [TSTPM21]. NeuralGIF addresses the challenge of animating clothed humans in unseen poses directly from raw 3D scans, thereby eliminating the need for manual registration or reliance on predefined mesh templates. Traditional mesh-based methods suffer from topological rigidity and limited expressiveness, particularly when modeling complex clothing such as open jackets, layered clothing, or clothing with significant non-rigid motion. NeuralGIF overcomes these limitations by proposing an animatable model of humans using neural implicits.

NeuralGIF is a generalized implicit function framework that models clothed human geometry using a pose-dependent backward skinning field coupled with a deformation field in canonical space. This architecture learns to map points from posed space to a canonical space and applies non-rigid deformations in that space before evaluating a signed distance function (SDF), enabling pose-conditioned shape reconstruction without any explicit mesh topology. By bridging ideas from classical forward skinning and neural implicit representations, NeuralGIF provides a principled and flexible formulation for modeling complex clothing dynamics. It significantly advances the state of the art in animating 3D human avatars and opens new directions for learning from raw, unregistered 3D scans.

To facilitate learning and evaluation of an animatable model of people in complex clothing, we also introduce ClothSeq, a novel 4D dataset of clothed humans, capturing real-world dynamic clothing behavior of highly complex clothing. The dataset includes sequences of subjects performing various motions in challenging clothing such as shrugs, puffers, and layered outfits. ClothSeq is the first dataset to provide dense scan sequences of such complex clothing, offering a valuable benchmark [QXL*22, CJS*23] for evaluating and training models that aim to handle dynamic and deformable clothing in motion.

1.2.3 TailorDiff: Unified and Expressive Generative Model for 3D Clothing

In **Chapter 6**, we propose TailorDiff, a novel SMPL-like foundation model for 3D clothing that is explicit, unified, and fully compatible with human body models like SMPL. Unlike prior works that require separate models for each clothing type or class or rely on mesh templates with fixed topology, TailorDiff uses a point cloud-based hierarchical latent diffusion model that is topologically agnostic and explicitly conditioned on body geometry. This enables the generation of diverse, well-aligned clothing across multiple styles, shapes, and fine-grained attributes (e.g., sleeve and hem length) using a single unified model.

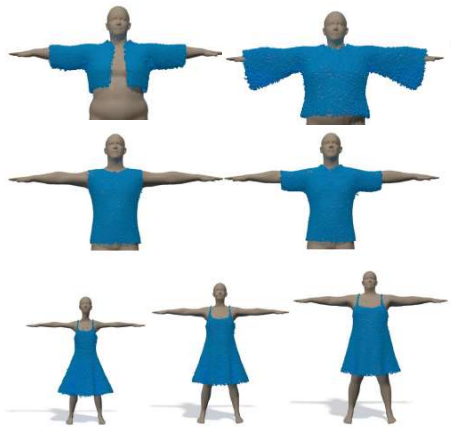


Fig. 1.3 **TailorDiff:** A topologically agnostic model for clothing generation, editing and inverse problems.

Another key contribution of this work is the curation and alignment of multiple large-scale synthetic datasets to create a pair of well-aligned clothing items and underlying body shape. This large-scale dataset is used to learn an expressive latent space of clothing conditioned on body shape, and subsequently, a DDPM is trained on the latent space. TailorDiff's latent space can be utilized for downstream tasks, such as reconstruction from partial observations and fitting to scans with a layered clothing representation. Prior methods can't handle diverse clothing topology, making them unsuitable for such tasks.

TailorDiff unites the expressive power of generative diffusion models with explicit point cloud representations, enabling layered clothing modeling and fine-grained semantic control, making it compatible with applications such as rendering, simulations that surpass the capabilities of prior approaches.

1.2.4 PoseNDF: Modeling Human Pose Manifolds with NDFs

In **Chapter 7**, we present "Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields" [TAL*22]. PoseNDF introduces a fundamentally novel formulation for modeling the manifold of plausible 3D human poses, departing from conventional approaches that rely on Gaussian assumptions or latent space regularization. Prior methods often model human pose using VAEs or GMMs, which inherently bias predictions toward the mean pose and restrict diversity, particularly in edge cases or under occlusion. In contrast, PoseNDF eliminates the need for any explicit latent space structure or prior distribution by directly learning a Neural Distance Field (NDF) over the SO(3) joint rotation space. This is the first work to leverage NDFs for learning data manifolds in such a high-dimensional domain, enabling the pose space to be defined implicitly as a continuous, differentiable function without any assumptions on manifold.

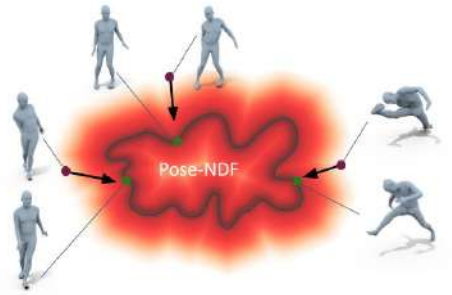


Fig. 1.4 **PoseNDF:** A NDF-based prior for modeling plausible human poses.

The method defines the zero-level set of the NDF as the manifold of valid human poses and uses analytic gradients of the learned field to project invalid or noisy poses onto this manifold. This results in a robust and interpretable mechanism for pose denoising, completion, and optimization-based tasks. Since the NDF provides a continuous, differentiable surface of valid poses, it can be readily used as a drop-in prior for optimization-based tasks, such as inverse kinematics, pose estimation from images, or scan registration. Unlike VAE-based models, it does not require sampling from a latent distribution, and its gradient-based projection makes it especially well-suited for use in differentiable pipelines. PoseNDF thus represents a powerful and general-purpose pose prior that not only improves accuracy and diversity in downstream applications but also sets a foundation for manifold learning using neural fields beyond Euclidean data. Recent works [DFZ*24, RLGCC23] have extended this beyond 3D human pose to motion and more generic datasets.

1.3 Publications and Works

Parts of the work presented in this thesis have previously been published in peer-reviewed venues. The following publications contain material that has been incorporated into this thesis and are listed in the order in which the corresponding content appears in the chapters above:

- [TBTPM20] **Garvita Tiwari**, Bharat Lal Bhatnagar, Tony Tung, Gerard Pons-Moll "SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing", European Conference on Computer Vision (ECCV) 2020, Oral.
- [TSTPM21] **Garvita Tiwari**, Nikolaos Sarafianos, Tony Tung, Gerard Pons-Moll "Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing", International Conference on Computer Vision (ICCV) 2021.
- [TAL*22] **Garvita Tiwari**, Dimitrije Antic, Jan Eric Lenssen, Nikolaos Sarafianos, Tony Tung, Gerard Pons-Moll "Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields", European Conference on Computer Vision (ECCV) 2022, Oral-Best Paper Honourable Mention.

This thesis also includes an unpublished work in which we revisit and explore the use of explicit representations for modeling clothing.

- [TXL*25] **Garvita Tiwari**, Yuxuan Xue, Jan Eric Lenssen, Tony Tung, Gerard Pons-Moll "TailorDiff: A Unified and Expressive Generative Model for 3D Clothing".

In addition, contributions were made to the following publications, which are, however, not part of this thesis:

- [ATO*24] Dimitrije Antić, **Garvita Tiwari**, Batuhan Ozcomlekci, Riccardo Marin, Gerard Pons-Moll "CloSe: A 3D Clothing Segmentation Dataset and Model", International Conference on 3D Vision (3DV), 2024.
- [HTB*24] Yannan He, **Garvita Tiwari**, Tolga Birdal, Jan Eric Lenssen, Gerard Pons-Moll "NRDF: Neural Riemannian Distance Fields for Learning Articulated Pose Priors", Conference on Computer Vision and Pattern Recognition (CVPR) 2024, Highlight, 11.9 % of accepted papers.
- [HTZ*25] Yannan He, **Garvita Tiwari**, Xiaohan Zhang, Pankaj Bora, Tolga Birdal, Jan Eric Lenssen, Gerard Pons-Moll "LiMo: Lifelike Human Motion Generation with Continuous-space Generative Models", Under Review.

1.4 Outline

In this section, we outline the thesis structure to guide the reader through the progression of the research conducted.

Chapter 2 Background: Following this introductory chapter, the thesis continues with a background chapter, which introduces key concepts necessary to understand the subsequent chapters. This includes an overview of linear blend skinning, a classical and widely used approach for animating or reposing mesh-based avatars. We then discuss a standard human body model, the SMPL model [LMR*15], which is used throughout all chapters in this thesis. We also describe commonly used 3D representations – such as meshes, point clouds, and neural implicit surfaces – which form the foundational representations used in the methods developed in the works presented here. In addition, we provide a brief introduction to generative models and diffusion models, both of which are especially relevant to the work presented in Chapter 6.

Chapter 3 Related Work: We present a comprehensive review of related works, focusing on both humans in clothing and human pose models. For human and clothing modeling, we trace the evolution of combined human-cloth representations as well as approaches that model clothing independently of the human body. One of the main focuses of this thesis is to explore various 3D representations for modeling humans and clothing. To reflect this, the related work chapter is organized to review previous research based on the types of 3D representations used. This structure helps highlight how different methods approach the problem and how they relate to the goals of the work presented in this thesis. We also provide a brief commentary on relevant concurrent work and research published after our contributions, to offer a broader perspective on the research landscape in the respective field. A similar structure is adopted in the section that discusses related work on human pose models.

The main body of this thesis is organized into two major parts, each reflecting one of the primary research directions explored during the course of doctoral studies.

Part I Human in Clothing: This part of the thesis is dedicated to the research of modeling humans in clothing, an area that has seen increasing attention due to its relevance in graphics, vision, AR/VR, content creation, and virtual try-on applications. This part comprises three chapters, two based on a peer-reviewed publication: SIZER [TBTPM20], NeuralGIF [TSTPM21], and one unpublished work, TailorDiff. Collectively, these works investigate different paradigms for representing humans and clothing, progressing from learning specific deformation models to building more expressive and unified clothing representations. Each chapter provides an in-depth discussion of the problem statement, specific 3D representation used, along with its advantages, limitations, and potential impact within the broader context of human-clothing models.

- **Chapter 4: SIZER** This chapter presents the first method for modeling size-sensitive clothing deformation in 3D. It introduces a new large-scale dataset featuring real humans wearing garments in multiple sizes and proposes two neural networks: ParserNet for semantic layer decomposition and SizerNet for deformation conditioned on body shape and clothing size. We detail the dataset acquisition and processing pipeline, as well as the architectures of both networks. In addition, we present baseline comparisons and ablation studies to demonstrate the effectiveness of our approach.

- **Chapter 5: NeuralGIF:** In this chapter, we introduce a neural generalized implicit function for animating clothed humans directly from raw 3D scans. Our approach leverages a backward skinning field along with a deformation field defined in canonical space to capture dynamic clothing behavior. We further extend the model to include a body shape-dependent deformation component. We show the model's effectiveness in capturing complex and highly dynamic deformations, and we compare its performance with state-of-the-art methods. Additionally, we provide an in-depth discussion of traditional skinning techniques and neural implicit skinning and deformation.

- **Chapter 6: TailorDiff:** We introduce TailorDiff, a topology-agnostic diffusion-based generative model for 3D clothing that is compatible with SMPL. Built on a point cloud VAE, TailorDiff learns a denoising diffusion probabilistic model (DDPM) in the latent space, conditioned on body shape. This enables the generation of diverse, well-aligned, and easily editable clothing. We provide details on data processing, model architecture, and training procedures. We demonstrate the effectiveness of our model across diverse generation and editing tasks, and showcase its utility in the downstream application of layered clothing fitting from scans, with comparisons to prior methods.

Part II Human Pose Priors: In this part, we shift the focus to human pose modeling, and present our work on the PoseNDF [TAL*22], which introduces a novel neural distance field-based representation for human poses.

- **Chapter 7: PoseNDF** PoseNDF introduces a novel approach to modeling the manifold of valid 3D human poses using Neural Distance Fields (NDFs), without relying on Gaussian priors or latent space structures. PoseNDF serves as a powerful, interpretable prior for various downstream tasks. We compare PoseNDF against existing pose priors in motion denoising, pose completion, interpolation, and pose estimation from images, showing an improvement in performance.

Chapter 8 Conclusion: This final chapter summarizes the key findings, contributions, and insights from the different parts of the thesis. We reflect back on the key findings and discuss promising directions for future research. We also provide a broader perspective on the topic of 3D digital humans, which is relevant to the current research landscape and application goals.

2

Background

2.1 Digital Avatars

Digital humans are widely used in various applications, including virtual reality (VR), augmented reality (AR), gaming, animation, telepresence, and interactive media. Typically, creating these avatars involves modeling the geometry, applying textures, and defining a suitable rig or skeleton to control movements. In this section, we cover two crucial concepts used for Digital humans and are prerequisites for most of the chapters in this thesis. The fundamental step in animating avatars realistically and efficiently involves determining how the character's geometry deforms in response to changes in the underlying skeletal structure. This process is commonly known as "rigging and skinning." A standard body model incorporates skinning and other deformations in the model itself. In this thesis, we focus on modeling human avatars, human avatars in clothing, and clothing using various representations. One of the key focuses of these works is making the shape reposable and animatable. As a result, both concepts of skinning and body models are frequently used in this thesis. Since the chapters in this thesis mainly build upon Linear Blend Skinning and SMPL body model, we first discuss LBS in Section 2.1.1 and SMPL in Section 2.1.2

2.1.1 Linear Blend Skinning

2.1.1.1 Rigging and Skinning

In 3D animation, rigging and skinning are the processes of binding a 3D model's surface (or mesh) to a skeletal structure (the rig) to allow for realistic movement and deformation during animation. The core idea of skinning is to link mesh vertices to skeleton bones, allowing for smooth and realistic animations. The skeleton, often represented by interconnected bones and joints, acts as the internal structure that defines the avatar's pose. Each vertex in the mesh moves according to its associated bones, typically influenced by multiple bones to allow smooth deformations and realistic articulation.

Standard skinning methods include Linear Blend Skinning (LBS), Dual Quaternion Skinning, and more advanced nonlinear approaches. Among these, Linear Blend Skinning remains the most popular due to its simplicity and computational efficiency, despite some limitations such as the "candy-wrapper" artifact at joints.

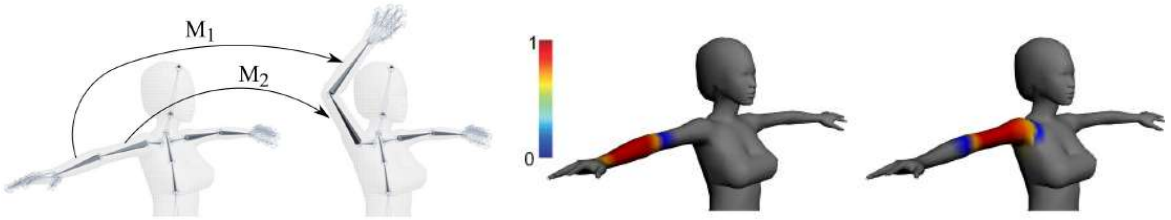


Fig. 2.1 **Linear Blend Skinning** [JDKL14]: *Left*: Bone transformations for a given deformed pose. *Right*: Blend weights corresponding to the lower and upper arm bones.

2.1.1.2 Linear Blend Skinning

Linear Blend Skinning, also known as Skeletal Subspace Deformation (SSD), is the standard and most straightforward technique for skinning. In LBS, each vertex of the mesh is assigned weights corresponding to multiple bones, determining the influence of each bone's transformation on the vertex. The final transformed vertex position is computed as a weighted linear combination of the transformations of influencing bones.

Formally, given a vertex $\mathbf{v}_i$ in rest position, the transformed position $\mathbf{v}'_i$ is computed as:

$$\mathbf{v}'_i = \sum_{j=1}^{B} (w_{ij}\mathbf{T}_j)\mathbf{v}_i \tag{2.1}$$

where:

- B is the total number of bones influencing the vertex.
- w_{ij} are the skinning weights indicating the influence of bone j on vertex i, with constraints $\sum_{j=1}^{B} w_{ij} = 1$ and $w_{ij} \geq 0$.
- $\mathbf{T}_j$ is the transformation matrix (usually consisting of rotation, translation, and scaling) of bone j.

Each transformation matrix $\mathbf{T}_j$ is calculated by combining the bone's local transformation relative to its parent bone along with the global transformations accumulated through the skeleton hierarchy. The transformation in Equation (2.1) is direct skinning, also known as forward skinning, as it involves transitioning from the canonical and rest poses to the deformed pose. We will revisit these terms and concepts in Chapter 5.

2.1.1.3 Limitations and Extensions of LBS

Despite its simplicity and widespread use, LBS has inherent limitations. One of the most common issues is the *candy-wrapper* artifact, where twisting rotations around bones cause the geometry to shrink or collapse unnaturally. This artifact results from linearly interpolating rotation matrices. Another limitation is the simplistic assumption that the skinning weights are pose-independent.

Several approaches have been proposed to mitigate these issues:

- **Dual Quaternion Skinning**: Uses dual quaternions for blending rotations and translations, significantly reducing artifacts common in LBS.
- **Nonlinear Skinning Methods**: Techniques such as spherical blend skinning and corrective blend shapes provide more accurate deformations.
- **Data-driven and machine learning-based methods**: Recently, neural network-based approaches have been proposed to learn more accurate skinning functions directly from data.

However, LBS remains a preferred starting point or baseline method due to its simplicity, efficiency, and compatibility with real-time applications. For a more detailed study on skinning and its impact, please refer to [JDKL14]. In Chapter 5, we dive deep into LBS, its limitations, and how to adapt LBS to neural implicit-based representation using volumetric skinning field and the formulation of backward skinning.

2.1.2 Body Models

Human body models provide structured, parameterized representations of human shapes and poses. These models allow efficient modeling and reconstruction of human geometry for animation, motion capture, and simulation. These models are efficient for rendering and solving inverse problems, and are compatible with standard graphics software and rendering engines. Prior to the introduction of works like [HLRB12, LMR*15], commercial methods predominantly involved manually rigging meshes and sculpting blend shapes to resolve traditional skinning issues, a labor-intensive and cumbersome process. Notable improvements in realism have been achieved over the past decade through the utilization of statistical human body models trained on data. In this section, we talk about the SMPL [LMR*15] body model, which has been used in all of the works [TBTPM20, TSTPM21, TAL*22, TXL*25] mentioned in this thesis.

2.1.2.1 SMPL: A Skinned Multi-Person Linear Model

The seminal work presented by Loper et al. [LMR*15] focuses on developing realistic human body models capable of accurately representing diverse body shapes and poses. Skinned Multi-Person Linear Model (SMPL) robustly captures a broad spectrum of body shapes and naturally exhibits shape and pose-dependent deformations. In Figure 2.3, we provide a visualization of the SMPL body model, where we see how different kinds of deformations are linearly added to the rest pose of the body model and finally articulated to the desired pose. It is designed to be differentiable and suitable for optimization-based and learning-based tasks. SMPL represents a human body mesh with a fixed topology and is parameterized by **pose** and **shape** parameters.

Mesh Representation: The SMPL model is represented by a triangulated mesh $\mathcal{M} = (\mathcal{V}, \mathcal{F})$ with $N = 6890$ vertices $\mathcal{V} \in \mathbb{R}^{N\times 3}$ and a fixed face connectivity $\mathcal{F}$. The output vertex positions depend on two sets of parameters:

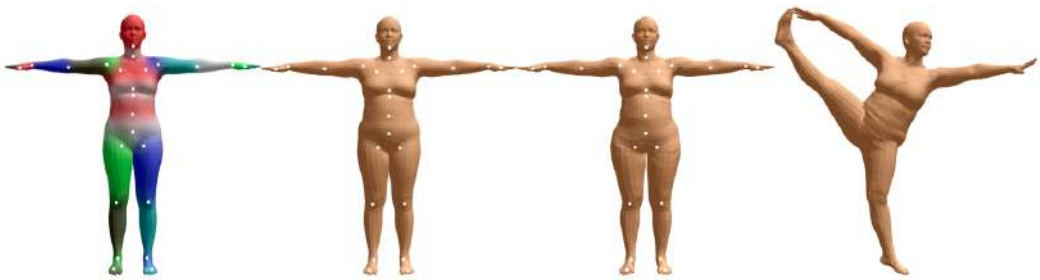


Fig. 2.2 **SMPL body model:** Template mesh with predefined blend weights, where shape and pose-dependent deformation are added(*middle*) and finally articulated to the given pose(*right*). (Courtesy of [LMR*15])

- **Shape parameters** $\beta \in \mathbb{R}^{|\beta|}$ (10 or 300 PCA coefficients)
- **Pose parameters** $\theta \in \mathbb{R}^{K\times 3}$ (axis-angle rotations for each of the $K = 24$ joints [LMR*15])

Latest SMPL-based models, such as SMPLX [PCG*19], have incorporated additional joints, including hand key points [RTB17] and facial features [LBB*17]. In this section, we will discuss only the SMPL body model, as 1) we have mainly used SMPL throughout our project, and 2) a similar formulation can be easily extended to SMPLX.

Shape Blend Shapes: Shape variation is captured using a learned blend shape basis $\mathbf{B}_S \in \mathbb{R}^{N\times 3\times|\beta|}$. The template mesh $\bar{\mathbf{T}} \in \mathbb{R}^{N\times 3}$ is deformed as:

$$\mathbf{T}_\beta = \bar{\mathbf{T}} + \sum_{i=1}^{|\beta|} \beta_i \mathbf{B}_{S,i}$$

Pose Blend Shapes: Pose-dependent deformations (e.g., muscle bulging) are modeled using pose blend shapes $\mathbf{B}_P \in \mathbb{R}^{N\times 3\times K\cdot 9}$, which depend on deviations of joint rotations from the rest pose:

$$\mathbf{R}_k = \text{Rodrigues}(\theta_k), \quad \forall k \in [1, K]$$
$$\mathbf{R}_{\text{pose}} = \text{vec}\left([\mathbf{R}_k - \mathbf{I}]_{k=1}^{K}\right)$$
$$\mathbf{T}_{\text{pose}} = \mathbf{T}_\beta + \mathbf{B}_P \mathbf{R}_{\text{pose}}$$

Linear Blend Skinning (LBS): The posed and shaped mesh $\mathcal{V}$ is obtained using Linear Blend Skinning (LBS):

$$\mathbf{v}_i = \sum_{j=1}^{K} w_{ij}\, \mathbf{G}_j(\theta, \mathbf{J}(\beta)) \cdot \mathbf{v}_i^*$$

where:

- $\mathbf{v}_i^* \in \mathbb{R}^4$ is the homogeneous coordinate of vertex i in the rest pose

- $w_{ij} \in [0,1]$ are blend weights of vertex i for joint j
- $\mathbf{G}_j(\theta, \mathbf{J}) \in \mathbb{R}^{4\times4}$ is the global rigid transformation of joint j, obtained from forward kinematics
- $\mathbf{J}(\beta)$ is the joint location regressed from the shaped template

The final posed mesh is given by Equation (2.2) and parameters $\bar{\mathbf{T}}, \mathcal{W}, S, \mathcal{J}, P$ are learned. We explain each parameter afterwards.

$$\mathcal{M}(\beta,\theta) = W(T_P(\beta,\theta;\bar{\mathbf{T}},S,P),J(\beta;\mathcal{J},\bar{\mathbf{T}},S),\theta,\mathcal{W}), \tag{2.2}$$

- Template mesh $\bar{\mathbf{T}}$: A mean template mesh $\bar{\mathbf{T}} \in \mathbb{R}^{3N}$ representing vertices in a neutral pose
- Shape blend shapes $\mathbf{B}_S$: Shape blend shapes capturing identity-dependent body variations.
- Pose blend shapes $\mathbf{B}_P$: Pose-dependent blend shapes accounting for pose-induced deformations.
- Joint regressor $\mathbf{J}(\beta)$ for accurate pose-dependent joint placements
- Skinning weights $\mathbf{W} \in \mathbb{R}^{N\times K}$: Blend weights for smooth vertex transformations.

Parameters $\bar{\mathbf{T}}, \mathcal{W}, S, \mathcal{J}, P$ are learned by minimizing reconstruction errors using aligned meshes from two datasets: a primary set consisting of 1786 registered meshes from 40 subjects and a multi-shape dataset from CAESAR. For detailed training procedures refer to the original publication [LMR*15].

2.2 3D Representations for Human in Clothing

Understanding and representing 3D geometry is a fundamental challenge in computer graphics, vision, and machine learning. Various 3D representations for modeling clothing and humans have been proposed, each with unique strengths and limitations. This chapter discusses key 3D representations: meshes, point clouds, and implicit functions, which are used throughout this thesis.

A powerful 3D representation is a critical component of a learning-based 3D content creation framework. An ideal 3D representation for high-quality reconstruction and synthesis should

- Capture fine-grained geometric details,
- Supports arbitrary topology,
- Be compact and memory-efficient, and
- Enable differentiable rendering and learning frameworks.
- Easily integrable with existing graphics and rendering pipeline for downstream applications.
- Compatible with learning frameworks

2.2.1 Meshes

Meshes are among the most widely used 3D representations in computer graphics and geometric processing. A mesh represents a surface $\mathcal{M}$ using a set of vertices $V = \{\mathbf{v}_i \in \mathbb{R}^3\}$, edges E, and faces F (typically triangles). The surface is approximated by piecewise planar patches, often leading to a representation:

$$\mathcal{M} = (V, F), \quad \mathbf{v}_i \in \mathbb{R}^3, \quad f_j \in \mathbb{N}^3$$

Meshes enable efficient rasterization-based rendering and support operations such as texture mapping, smoothing, and deformation. Meshes can accurately model complex geometries and support operations like texture mapping, subdivision, surface editing, and regularization. They are widely supported by engines like Blender, Unity, and Unreal Engine. Moreover, body models [LMR*15, ASK*05] and clothing models [GCS*19, GRH*12] have demonstrated impressive results with mesh-based representations.

However, they have limitations:

- Poor adaptability to topological changes (e.g., merging holes or dynamic cloth).
- High-resolution meshes can be memory-intensive.
- Mesh reconstruction from raw data (e.g., point clouds) is challenging and often requires post-processing techniques such as Poisson reconstruction [KBH06].

2.2.2 Point Clouds

Point clouds represent 3D shapes as an unordered set of discrete points $P = \{\mathbf{p}_i\}_{i=1}^{N}$, where each $\mathbf{p}_i \in \mathbb{R}^3$ may carry additional features such as RGB color or normals. They provide a natural representation of raw sensor data acquired via LiDAR, multi-view stereo, or depth cameras.

$$P = \{\mathbf{p}_i \in \mathbb{R}^3\}, \quad \mathbf{p}_i = (x_i, y_i, z_i)$$

Point clouds are highly flexible and topologically agonistic but suffer from several challenges:

- Lack of explicit connectivity impedes operations like surface reconstruction.
- Sensitive to noise, non-uniform sampling, particularly in low-density regions.
- Lack of underlying structure also makes it difficult to introduce any surface regularizer, such as Laplacians.

Despite these drawbacks, point clouds are widely used in autonomous navigation, robotic manipulation, and SLAM due to their simplicity and flexibility. Moreover, the flexibility and topologically agonistic nature of point clouds make them suitable for learning-based tasks. Pointcloud-based methods, such as PointNet [QSMG17] or PointNet++ [QYSG17a], are widely used for classification, segmentation, and generation tasks. Moreover, recent 3D Gaussian Splatting [KKLD23] can also be driven using a pointcloud-based representation.

2.2.3 Implicit Representations

Neural implicit representations have recently emerged as powerful methods for representing and reconstructing 3D shapes using neural networks [XTS*22]. These methods utilize continuous implicit functions that define geometric surfaces through level sets, typically relying on signed distance fields (SDFs) or occupancy functions.

2.2.3.1 Signed Distance Functions and Occupancy Fields

Implicit representations encode 3D geometry as a continuous function defined over space, such as signed distance functions (SDF) or occupancy fields. These methods describe whether a point lies inside or outside a surface or provide the signed distance to the nearest surface.

Classically, a 3D shape can be implicitly represented using a Signed Distance Function (SDF), which for any point $\mathbf{x} \in \mathbb{R}^3$, defines the distance from $\mathbf{x}$ to the nearest surface, with a sign indicating whether the point is inside or outside the object:

$$\mathrm{SDF}(\mathbf{x}) = \begin{cases} -d(\mathbf{x}, \partial\mathcal{M}) & \text{if } \mathbf{x} \text{ is inside the shape } \mathcal{M}, \\ 0 & \text{if } \mathbf{x} \text{ is on the surface } \partial\mathcal{M}, \\ +d(\mathbf{x}, \partial\mathcal{M}) & \text{if } \mathbf{x} \text{ is outside the shape } \mathcal{M}, \end{cases}$$

where $d(\mathbf{x}, \partial\mathcal{M})$ denotes the Euclidean distance to the shape boundary $\partial\mathcal{M}$.

Alternatively, the occupancy function formulation defines shapes by indicating whether points are inside or outside:

$$\mathrm{Occ}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \text{ is inside the shape,} \\ 0 & \text{otherwise.} \end{cases}$$

2.2.3.2 Neural SDFs, Neural Distance Fields, and Variants

Recent advancements combine SDFs with neural networks to learn implicit surface representations from raw data. Methods like DeepSDF [PFS*19a] learns signed distance fields, OccNet [MON*19a], IF-net [CAPM20] learn occupancy given global or hierarchical features, and Neural Distance Fields (NDF) [CPM20] learn an unsigned distance field for open surfaces. NeRF [MST*20] extends implicit formulations to encode both geometry and appearance implicitly by modeling a 3D volume's radiance and density. NeUS [WHH*23] integrates SDFs and NeRF for better geometry.

These neural implicit representations significantly advance capabilities for representing complex 3D structures and visual content, enabling high-fidelity reconstruction, shape generation, and rendering from sparse or incomplete observations. Neural implicits are differentiable and allow for easy integration into learning frameworks for tasks such as shape optimization, novel view synthesis, and are capable of detailed and accurate 3D reconstruction from sparse data. However, the inference can be slow due to dense sampling requirements. Moreover, it's not straightforward to represent open surface meshes using

this formulation. Some methods [GSF22] have introduced a formulation to represent shapes using UDFs and reconstruct meshes using modified marching cubes.

2.3 Generative Models

In the later chapters of this thesis, we enter the domain of generative models. In Chapter 6 we utilize state-of-the-art diffusion models for learning a unified clothing model. In Chapter 7, we propose a novel formulation of learning a data manifold, which is closely tied to generative models. We discuss a diffusion-based generative model in this section, as it will help the reader understand the TailorDiff model and provide interesting insights into the PoseNDF formulation for learning the data manifold, as we discuss later in Section 7.3. In this section, we first discuss diffusion models in Section 2.3.1, followed by a background of latent diffusion model in Section 2.3.2. In Chapter 6, we use the insights from latent diffusion models to learn a point cloud diffusion model.

2.3.1 Diffusion Models

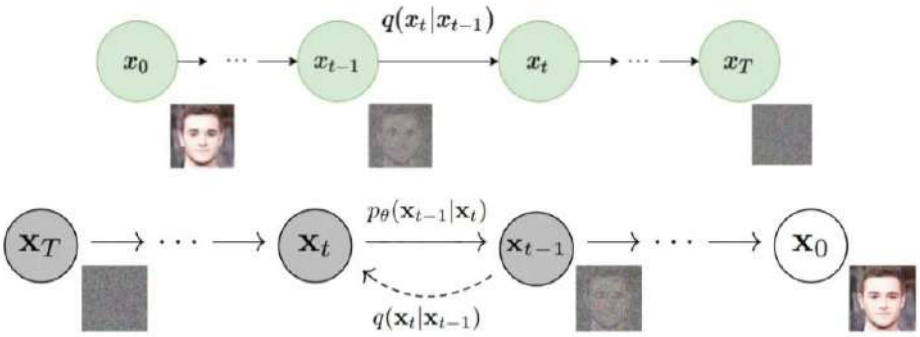


Fig. 2.3 **Diffusion Models:** *Top:* Introducing noise via Forward diffusion process. *Bottom:* Modeling the reverse process with a neural network.

Diffusion models have emerged as powerful generative frameworks capable of synthesizing high-quality data across diverse modalities, including images [RBL*21], video [WAW*25], and 3D [ZDW21, CYAE*20]. At their core, diffusion models represent data generation as a sequential denoising process, starting with random noise and iteratively reducing this noise to reconstruct samples from the underlying data distribution. Diffusion models rely on two main processes: a forward diffusion process and a reverse denoising (generative) process.

Forward Diffusion Process The forward diffusion process gradually adds Gaussian noise to an original data sample x_0 over T discrete timesteps, resulting in increasingly noisy intermediate samples x_t, until the final timestep x_T is nearly pure Gaussian noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

The joint distribution over all timesteps can be expressed as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

where β_t is the variance schedule controlling the noise level at timestep t.

Reverse Denoising Process The reverse process aims to gradually remove noise and recover the original data. It is modeled by a parameterized neural network θ:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$$

The neural network learns the conditional distribution to predict the mean $\mu_\theta(x_t,t)$ and variance $\Sigma_\theta(x_t,t)$ that best approximate the true reverse conditional distribution.

In practice, the forward diffusion process is straightforward and requires no training since it simply adds Gaussian noise according to a predefined schedule. The reverse denoising process is where training occurs. The network is trained to minimize the difference between its predictions and the true data samples at each timestep. This is typically achieved by optimizing the network to predict the noise added at each time step, thereby reducing the complexity of training.

During generation, the process begins by sampling a pure Gaussian noise x_T. The trained network then progressively denoises this noisy input over T timesteps, gradually shaping the noise into a structured data representative of the learned data distribution.

One of the key strengths of diffusion models is their ability to produce highly realistic and diverse samples, capturing intricate details while allowing substantial control through conditioning mechanisms. Diffusion models outperform earlier generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), in terms of training stability and the quality of generated samples. Diffusion models have numerous applications, including image synthesis, text-to-image generation, audio waveform synthesis, and video generation. Their versatility and robustness enable significant advancements in conditional generation tasks, providing user-guided and context-aware outputs across various domains.

2.3.2 Latent Diffusion Models

Latent Diffusion Models (LDMs) are a computationally efficient variant of Denoising Diffusion Probabilistic Models (DDPMs) that operate in a learned latent space rather than the original high-dimensional data space. The key idea is to decouple the compression and generative processes by first learning a perceptually meaningful latent space through an autoencoding architecture, and then applying the diffusion process in this compressed space.

Given an input image $x \in \mathbb{R}^{H\times W\times C}$, an encoder $\mathcal{E}$ maps it to a latent code $z = \mathcal{E}(x) \in \mathbb{R}^{h\times w\times c}$, where typically $h < H$, $w < W$, and $c \geq C$. The decoder $\mathcal{D}$ is trained to reconstruct the image as $\hat{x} = \mathcal{D}(z)$. Once trained, the diffusion model is applied to the latent variable z.

In the diffusion process, Gaussian noise is progressively added to the latent code over T timesteps using a Markov chain:

$$q(z_t \mid z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\, z_{t-1}, \beta_t \mathbf{I}),$$

where β_t denotes the variance schedule. The model is trained to learn the reverse process $p_\theta(z_{t-1} \mid z_t)$ using a neural network ε_θ that predicts the added noise at each timestep. The training objective typically minimizes a reweighted form of the variational bound:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{z_0,\varepsilon,t}\left[\|\varepsilon - \varepsilon_\theta(z_t, t)\|^2\right],$$

where $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$.

At inference time, sampling begins from Gaussian noise $z_T \sim \mathcal{N}(0, \mathbf{I})$, and the denoising process is iteratively applied to obtain z_0, which is then decoded by $\mathcal{D}$ to generate the final image.

By operating in the latent space, LDMs significantly reduce memory and computational demands, enabling training and inference with high-resolution images. Moreover, conditioning mechanisms (e.g., CLIP embeddings, segmentation maps, text prompts) can be easily integrated into the denoising network, making LDMs highly versatile for diverse tasks such as text-to-image synthesis, inpainting, and super-resolution.

3

Related Work

In this chapter, we discuss related works on *Human in Clothing*, *Human Pose Priors*, and closely related fields. We begin with *Human in Clothing* in Section 3.1, where we first summarize the history of clothing models, including earlier physics-driven and hand-crafted approaches. We then provide insights into datasets relevant to human modeling and further discuss human and clothing models developed to address different goals and based on various representations. For each section, we also review concurrent and follow-up research works to provide a broader view of the research landscape in the respective field. For *Human Pose Modelling* in Section 3.2, we first discuss previous works on pose and motion priors, followed by a discussion on using neural implicit representations to learn the data manifold.

3.1 Human in Clothing

In this section, we review relevant work for the first part of the thesis, Human in Clothing. We first give a brief overview of more classical and traditional approaches to modeling digital humans. Throughout this thesis, we explore various representations for human and clothing models, aiming to address a range of goals – from learning from structured data to developing topology-free models, and ultimately propose a unified model. Accordingly, we organize the related works to reflect these objectives. We also provide a detailed overview of 3D human and clothing datasets discussed in this chapter, which are referenced in subsequent chapters.

3.1.1 Digital Avatars and Clothing: A Historical Overview

3.1.1.1 Physics-based Modeling of Human and Clothing

Classical virtual human modeling and clothing modeling heavily relied on physics-based simulation for designing and deformation. The core concepts behind these works are detecting and solving self-collisions [VT94]. However, the high cost, limited usability, incompatibility with learning frameworks, and difficulty in modeling clothing and deformable objects under forces such as motion and friction make these methods challenging to use and scale for real-time applications. Some initial work [VCMT95] presented efficient techniques for simulating cloth as a deformable surface, laying the groundwork for stable cloth animation in computer graphics. [BW98] enabled fast yet stable simulation by the use of implicit integration with stiff springs, allowing cloth simulators to take large time steps without blowing up. These methods primarily focused on large-scale deformation but lacked the ability to capture fine wrinkles. As highlighted in [HBVT99], *Wrinkles add life to garments in fashion* [HBVT99].

This work describes a method to simulate realistic wrinkles on clothes without requiring a fine mesh and minimizing large computational overhead by introducing an area conservation property, thereby making the large-step computation stable. Bending stiffness models were refined as well, e.g. [CK02] introduced an energy term for bending that captured fine creases and wrinkles. By the mid-2000s, physics-based cloth simulation had matured to the point of widespread adoption in industry. Film productions were routinely using these methods, a feat made possible by advances in cloth simulation research. Real-time applications also benefited: physics engines for games (e.g., Havok Cloth, NVIDIA PhysX) incorporated research findings, such as position-based dynamics, to simulate cloth in interactive settings.

In digital avatar modeling, physics-based strategies were explored mainly through anatomical simulation. Early work by [SPCM97] introduced anatomy-based modeling of human muscle and skin for animation. Such biomechanical models could produce very realistic muscle bulging and skin deformation by simulating underlying forces, but they required extensive manual modeling of bones and muscles and were computationally expensive. The following works [DCKY02] used finite-element simulations of soft tissue and muscle to achieve realism. While these physics-driven human models provided accuracy (important in medical or high-end graphics contexts), their complexity limited their use.

In most graphics applications, a simpler kinematic paradigm was favored: an articulated skeleton with skin deformed via weights (so-called skinning). Traditional skeletal animation with techniques like linear blend skinning [JDKL14], Spherical skinning [Kv05], or Eigen [KJP02] became standard in both research and industry. Skinning methods are fast and easy to control, but because the deformation weights were usually artist-defined rather than learned from real data [BP07], they often failed to reproduce subtle muscle dynamics or soft-tissue movement. Some work introduces dynamic effect using pose-dependent deformation [LCF00]. In practice, high-end character pipelines [MCC11] combined both approaches: artists would rig models with skeletons and sometimes add handcrafted muscle simulators or secondary cloth simulators for skin and clothing.

3.1.1.2 Data-Driven Models of Human Bodies and Clothing

As 3D scanning [3dM] and 3D/4D datasets [HSS*09, BRLB14, RBD*02] of human shape became available, data-driven modeling approaches gained traction. Instead of assuming hand-tuned physical parameters, these methods learn statistical representations of human bodies. In the context of human body avatars, an early milestone was the work [ACP03] that registered a common mesh template to 250 body scans of different people in identical poses. By applying PCA to the aligned scan data, they obtained a low-dimensional morphable model [EST*20] that could parameterize diverse body types. This was one of the first data-driven human shape models, enabling applications like shape completion and body shape estimation from sparse observations. This concept is still widely used in current body models [LMR*15, ASK*05] and clothing models [PLPM20, BTTPM19]. Another direction of work [ST03], used body measurement parameters such as shoulder length, circumference *etc*. to drive avatars, focusing on bridging the gap between anthropometric data and graphics models. Some work [ACPH06] also incorporates pose-dependent deformation from data by scanning a single person in multiple poses and then interpolating between the captured

poses to produce new poses for the given subject. However, this was subject-specific. One of the initial unified models of body shape and pose was SCAPE [ASK*05]. SCAPE introduced a learned model that factorized deformations into two subspaces: one for different identities (body shapes) and one for different poses. The SCAPE model represented pose-dependent deformations in a pose-normalized space (decoupling rigid bone rotation from non-rigid flesh deformation) and combined them with a PCA shape space for body variability. This was a seminal data-driven avatar model in graphics, and its influence is clear in later models [HSS*09, HLRB12, LMR*15]. Such linear and unified models were a precursor to the now-ubiquitous SMPL [LMR*15, PCG*19], which builds upon these works. We provide an overview of SMPL in Section 2.1.2.

In parallel, data-driven clothing models emerged as researchers sought to capture cloth draping behavior from real examples or high-resolution simulations. One thread of work aimed to leverage real-world data (images or video of clothing) to understand cloth draping on different bodies and poses [SGdA*10, RASS24, dAST*08, BTH*03, HX*19, YFHWW18]. Other works went further by bypassing physical simulation at runtime and learning from offline PBS data for animation. This remains an active area of research [GTBH23, PLPM20, VSGC20, SOC19, ZHW23]. We provide further discussion in this direction in the upcoming sections.

In summary, the historical evolution before 2015 – from pure physics-based simulation to hybrid and data-driven models – established the techniques and even the evaluation benchmarks that guided the next generation of deep learning models for digital avatars and clothing. Each seminal contribution, whether a new physical model or a learned statistical model, incrementally pushed the fidelity and functionality of virtual humans, directly shaping the advancements that followed.

3.1.2 Human in Clothing Datasets

We discuss relevant datasets for human and clothing models in this section. We divide this section into 3D real-world datasets, 4D and multiview video real-world datasets, and 3D and 4D synthetic datasets. For each subsection, we first discuss the datasets released prior to our publications and then provide a brief discussion of the datasets that followed our publications.

3.1.2.1 Real World 3D Datasets

Understanding humans in clothing requires large-scale, real-world datasets that capture the diversity of human bodies, poses, appearances, and clothing styles. PeopleSnapshot [AMX*18b] provides 3D reconstructions of people rotating in a fixed A pose and in front of a camera. It includes corresponding SMPL parameters and consists of 24 subjects in a fixed, simple A-pose. DeepHuman [ZYW*19] contains around 7,000 models spanning variations in pose, subject identity, and clothing. It was captured using the DoubleFusion [YZG*18] framework. MGN [BTTPM19] offers 3D scans of approximately 300 clothed subjects, aligned to the SMPL body model. These scans are captured via multi-view photogrammetry, resulting in higher-quality geometry compared to previous RGB or RGB-D based setups [SGdA*10].

Dataset	#Subjects	Registration	Loose Garments	Accessories	Segmentation
SIZER [TBTPM20]	100	SMPL+D,G	✗	✗	✓
CustomHuman [HILJO23]	643	SMPL-X	✗	✗	✗
THuman2.0 [YZG*21]	500	SMPL-X	✗	✗	✗
IIIT 3D Human [JSP*23]	150	SMPL	✓	✗	✗
2K2K [HPY*23]	2000	SMPL-X	✓	✓	✗
CloSeD [ATO*24]	100	SMPL+D	✗	✗	✓
BuFF [ZPBP17]	6	SMPL	✗	✗	✗
CAPE [MYR*20]	15	SMPL+D	✗	✗	✗
ClothSeq [TSTPM21]	3	SMPL	✓	✗	✗
ActorsHQ [IRG*23a]	8	✗	✓	✗	✗
X-Humans [SGK*23]	20	SMPL-X	✗	✗	✗
4DHumanOutfit [ABB*23]	14	SMPL	✓	✗	✗
4D-Dress [WHG*24]	64	SMPL-X	✓	✗	✓

Table 3.1 **Real-world Human in Clothing Datasets:** We present an overview of 3D(*top*) and 4D(*bottom*) scans of humans in clothing, comparing the datasets in terms of their variations and the types of annotations they provide.

Several other works rely on commercially available high-fidelity 3D scan datasets, such as RenderPeople [ren], Axyz [axy], Twindom [twi], and Treedy [tre], which provide thousands of textured 3D scans with variations in pose, subject, and clothing.

In this direction, we introduce SIZER [TBTPM20], a large-scale dataset of humans in clothing that captures variation in style, size, and body shape. More details are provided in Section 4.3. SIZER was the first large-scale dataset focusing on clothing size variation, containing 100 different subjects wearing casual clothing in multiple sizes, totaling approximately 2,000 scans. The dataset includes 3D scans, SMPL registrations, clothing part segmentations, garment category, and size labels. Compared to prior and follow-up works, SIZER includes a more demographically diverse set of participants—spanning Asian, South Asian, and Caucasian subjects—by inviting volunteers from an international community. In contrast, most existing datasets are limited in demographic representation due to their geographically constrained data collection setups. Today, SIZER is widely used for benchmarking tasks such as 3D human reconstruction [XXMPM24], 3D clothing segmentation [MMC23, ATO*24, SDK*25], and layered modeling of humans and clothing [CPA*21a, DLLG*23, LGF23].

Since 2020, the scale and realism of real-world 3D human-clothing datasets have increased significantly. THuman2.0 [YZG*21] and THuman2.5 provide high-quality, clean, textured 3D scans captured using a dense DSLR rig, consisting of 500 and 2.5k scans covering clothing, pose, and body shape variation. THuman3.0 [SYWL23] contains 20 human-clothing combinations, each with 15–35 high-quality scans. However, these datasets [YZG*21, SYWL23] and their follow-ups [ZHY*22, SZZ*22, ZSZ*21] are primarily limited to Asian subjects. IIIT 3D Human [JSP*23] comprises over 1,50 scans captured using a hand-held scanner, predominantly featuring Indian subjects in both traditional Indian and casual Western clothing. CustomHuman [HILJO23] includes over 643 personalized human scans, covering approximately 125 clothing items across 81 subjects. All these

datasets [YZG*21, HILJO23, TBTPM20] typically provide SMPL and/or SMPL-X fittings and are widely used for 3D reconstruction and animation tasks. Deepfashion3D [ZCJ*20] provides a high-quality real-world capture of common 3D clothing items.

2K2K [HPY*23] is among the largest datasets, comprising 2,000 identities, and offers a rich resource for learning generalizable clothing representations. However, many scans include accessories—from small items like phones to large objects like chairs—which introduce significant occlusions and limit the dataset's usability for pure human-clothing tasks. Additionally, a substantial portion of the dataset consists of scans of children. CloSe [ATO*24], built on top of SIZER, expands the dataset with more diverse poses and finer segmentation labels.

In summary, these datasets span from 100 to 2,000 scans covering different aspects of humans in clothing, such as pose-dependent deformation, clothing size variation, clothing draping on different bodies *etc*. This reflects a growing trend toward increased realism, subject diversity, and data scale to support modern learning-based approaches in clothed human modeling and reconstruction.

3.1.2.2 Real World 4D Datasets

Recent years have seen a surge in 4D human scan datasets aimed at capturing clothing dynamics, but their quality, realism, and diversity vary significantly. Early datasets such as ClothCap [PMPHB17a], Dyna [PMRMB15], and BUFF [ZPBPM17b] focus largely on minimal or simple clothing. ClothCap [PMPHB17a] introduced an approach to register dynamic cloth deformation onto body templates using sequences with minimal or tight-fitting clothing. Similarly, Dyna [PMRMB15, BRPMB17a] features 4D scans of 10 subjects, capturing tight sportswear with main focus on dynamic skin deformation, while BUFF [ZPBPM17b] captures only 6 subjects performing short sequences with limited clothing variation. These datasets provide valuable early benchmarks for clothing deformation but fall short in modeling highly complex garments. CAPE [MYR*20] increases the scale and diversity of the BUFF dataset by introducing more varied garments (e.g., Coats) across 15 subjects. However, it still primarily focuses on moderately loose clothing.

To this end, ClothSeq, a dataset proposed in one of our works [TSTPM21], marks a significant shift by providing real 4D scans of loose, everyday clothing, such as shrugs and puffer jackets (500–1000 frames per subject), which emphasizes high clothing complexity and dynamic deformation. Although ClothSeq includes only 3 subjects, it presents a more realistic and challenging benchmark for dynamic clothing modeling.

Building on this trend, 4D-DRESS [WHG*24] is a recent dataset of 4D scans of diverse, loose-fitting garments, including dresses, shirts, and open blouses. Its coverage of complex garments and a large subject pool make it particularly suitable for learning generalized clothing models. Similarly, 4DHumanOutfit [ABB*23] focuses on motion-conditioned outfit sequences, capturing 11 distinct motions per outfit across a wide variety of clothing styles, including dresses and jackets. While it lacks very long sequences, it provides important supervision for motion-aware generative modeling. ActorsHQ [IRG*23a] is one of the highest-quality 4D scans of actors in diverse outfits; however, it lacks texture in meshes and consists of only 8 subjects.

Dataset	Body Shape Variation	Clothing Categories	SMPL	Loose Clothing	Accessories	# Clothing Items
TailorNet [PLPM20]	✓	4	✓	✗	✗	∼ 600
DeepFashion3D [HYH*20]	✓	10	✓	✗	✗	563
CLOTH3D [BME20]	✓	7	✓	✗	✗	8.5k
3D Garments [KL21]	✗	19	✗	✓	✗	20k
ClothesNet [ZZL*23]	✗	–	✗	✗	✓	2.5k
GarmentCodeData [KKK*24]	✓	19	✗	✓	✗	115k
CLOTH4D [ZHW23]	✓	–	✓	✗	✗	1k
D-Layers [SLD23]	✓	10	✓	✓	✗	9.8k
TailorDiff [TXL*25]	✓	19	✓	✓	✗	∼ 100k

Table 3.2 **Synthetic Clothing Datasets:** Synthetic datasets consist of 3D Clothing simulated on different body shapes and poses and consist of 3D static scans (*top*) or dynamic 4D sequences (*middle*). Some recent datasets also provide loose clothing in their dataset. **TailorDiff**(*bottom*) combines diverse datasets by aligning meshes to a common frame and registering them to the SMPL model.

On the other hand, datasets like Thuman 3.0 [SYWL23] and X-Humans [SGK*23] offer high-quality static or pseudo-dynamic data with clothing that mostly conforms to body topology. Likewise, AvatarRex [ZZZ*23] employs a multiview setup and utilizes neural implicits [WHH*23, MST*20, LZWL24] to obtain a proxy 3D representation from a multi-view capture setup. The same applies to DyMVHumans [ZLL*24, XLL*23], which uses NeUS2 [WHH*23] to reconstruct high-quality motion from limited views; although visually impressive, the raw scans are not released.

In conclusion, while early datasets provided foundational tools for studying clothed human shapes, modern datasets, such as ClothSeq [TSTPM21], 4D-DRESS [WHG*24], 4DHumanOutfit [ABB*23], and ActorsHQ [IRG*23a], represent a new generation of high-fidelity 4D human datasets with diverse clothing, body types, and motions. We provide an overview of real-world 3D and 4D datasets in Table 3.1 and more detailed discussions in Section 4.3 and Section 5.5.

3.1.2.3 Synthetic Datasets

On the other hand, for separate clothing models, synthetic datasets have predominantly been used for learning individual clothing models due to the challenges in extracting clothing from real-world scans. Some works have addressed the problem of extracting clothing from real-world data, but they are either limited to predefined classes [TBTPM20], require carefully controlled data capture setups [LCT18b], or rely on dense capture of clothing [RGW*25]. Segmentation-based models for extracting clothing from real-world scans often result in partial observations [ATO*24, SDK*25].

Synthetic data, on the other hand, consists of 3D models of subjects with segmented clothing. 3DPeople [PSC*19] and Cloth3D [BME19] are large datasets of synthetic 3D humans with clothing. We provide an overview of datasets in Table 6.2. Some works provide offline PBS-based clothing simulation data to learn pose-, shape-, or sketch-dependent

clothing deformations [PLPM20, SOC19, WSFM, GTBH23]. In [JZH*20], the authors propose a mixture of synthetic and real data, which includes variations in garment, body shape, and pose. However, the real dataset portion (300 scans) is fairly small.

More recently, the scale and variation of 3D clothing datasets have increased tremendously. GarmentCodeData [KL21, KSH23] contains 115,000 data points that cover a variety of designs across many common garment categories—tops, shirts, dresses, jumpsuits, skirts, pants, etc.—fitted to a range of body shapes sampled from a custom statistical body model and rendered with three different textile materials. This dataset also provides 2D sewing patterns. D-LAYERS [SLD23] contains 700K frames capturing the dynamics of 4,900 different combinations of multi-layered garments, driven by both human motion and randomly sampled wind. ClothesNet [ZZL*23] consists of artist-designed clothing items and accessories, without an underlying body. Cloth4D [ZHW23] includes artist-designed clothing items created using 2D sewing patterns and provides 4D sequences of clothing simulated across different body shapes and motion sequences. In our latest work, we leverage large-scale synthetic clothing datasets to train a unified generative model for clothing. We combine multiple datasets, align the clothing meshes, and register SMPL body models to all clothing items where they are missing. This process results in a diverse and comprehensive collection of synthetic garments from various sources. Using this dataset, we train TailorDiff, our large-scale generative model for clothing.

3.1.3 Mesh-Based Approaches

In this section, we review existing approaches that use mesh-based models for a wide range of tasks involving humans and clothing. These include modeling non-rigid deformations, performing 3D and 4D reconstruction, and enabling shape and semantic understanding. Our focus spans both human body and clothing representations, highlighting methods that leverage mesh structures to capture geometry, dynamics, and parsing of clothing.

Accurately reconstructing 3D clothing with fine-grained details, such as wrinkles, is essential for achieving realism; however, it remains a challenging task. In this section, we provide an overview of mesh-based methods for representing humans in clothing, with a focus on modeling 3D clothing deformations, reconstructing clothed humans from images or videos, estimating body shape under clothing, and clothing parsing.

3.1.3.1 Physics-Based Simulation (PBS) for Clothing:

Physics-based simulation (PBS) techniques have been employed to model clothing dynamics [WHRO10, WRO11, MBT*12, GRH*12, Stu18, TZZ*19, PLPM20, SOC19, GCS*19]. 3D meshes are animated using PBS governed by some fundamental forces and collision detection. PBS methods require designing a garment/mesh template and manual intervention for finding suitable physical parameters. Past work has explored ways to automate this limitation by trying to infer these physical parameters mechanically from data [MBT*12, WRO11], or use videos to infer them [RKP*07, SGdA*10]. PBS methods are also computationally expensive, and the complexity increases for realistic animations. Several approaches generate realistic and high-frequency details by computing a high-resolution mesh from a base coarse mesh

using constraint-based optimization [MC10, RPC*10]. While simulations generate clean, noise-free data, they are computationally expensive, often require manual parameter tuning (e.g., time step, material properties), and can be unstable. Some approaches also use linear transformations between pose-shape and surface deformations [ASK*05, LMR*15, OBB20] to animate new poses and shapes.

3.1.3.2 Data-Driven Clothing Models:

In contrast to PBS, data-driven methods learn garment deformations directly from data, either generated via PBS or captured from real humans in clothing. With the advances in deep learning and 3D data acquisition [TNM08, TNM09, MNTT12], data-driven models of clothing and humans [XZT19] are becoming increasingly popular. These models are generally faster at inference and offer more flexibility for downstream tasks.

Modeling Human in Clothing from Images/Videos: Mesh-based explicit clothing representation has also been widely used for 3D reconstruction from images,video, or depth. This task differs from human pose and shape estimation tasks [KBJM18, CPB*20, KAB20], as the reconstructed mesh also contains details such as facial and clothing features. Early works using multi-view stereo reconstruct global body shape but miss fine wrinkles in textureless regions [WCF07, SH07, BPS*08, dAST*08, TNM09, LFB17]. A key breakthrough was made by [AMX*18b, AMX*18a], which reconstructed detailed clothed humans from monocular videos using SMPL with a displacement field. This led to numerous follow-ups [HX*19, AMB*19, APMTM19, SHN*19, HX*20, HXL*20], where the displacement field was learned. Although most do not explicitly separate body and clothing.

Modeling Clothing from Images/Videos: To better model loose garments and layered interactions, some works isolate clothing as a separate mesh layer over the body. Works like [BTTPM19, JZH*20, DMQ*23] represent humans and clothing as separate layers in the reconstruction. These methods often rely on coarse-resolution garment templates, with high-frequency wrinkles encoded in UV space using displacement, normal, or texture maps [LCT18a, APMTM19, LIPM19a]. However, the diversity of clothing represented by a pre-defined template is often limited, and we provide a detailed discussion of how to tackle this problem using different representations in Section 3.1.6.

Pose-Dependent Clothing Models: One key aspect in modeling humans and clothing is articulation and pose-dependent non-rigid deformations that follow body models. Since body models do not model personalized and clothing details, various methods [APMTM19, BTTPM19, PMPHB17b, JZH*20] have explored ways to extend the underlying body model, such as SMPL [LMR*15], by predicting per-vertex displacements for personalized and clothing features. However, they do not account for pose-dependent deformations. Many models account for deformations caused solely by pose changes [LCT18a, ZWCM20], often using template-specific fitting. Dyna [PMRMB15] models soft-tissue dynamics due to motion using a second-order autoregressive model.

Pose, and Shape-Dependent Clothing Models: Pose and shape-conditioned models offer more expressive clothing behavior. Several works learn such models from PBS-generated datasets [GRH*12, SOC19, PLPM20, GCS*19, BME19], capturing deformations across varying shapes and poses. More recent work, such as SNUG [SOC22] and Vidaurre et al.[VSGC20], focuses on learning dynamic deformation for a fixed template in a self-supervised manner, while others like [GTBH23] attempt template-agnostic deformation prediction.

Pose, Shape, and Style-Dependent Clothing Models: Going a step further, some models incorporate garment style in addition to pose and body shape [PLPM20, WCPM18, TBTPM20]. Generative models trained on real scans [MYR*20] can modify clothing style, body shape, and body pose to create holistic human-in-clothing images. Approaches such as [GRH*12, SOC19, GCS*19, PLPM20] utilize PBS to generate clothing deformation datasets for various body shapes, poses, or styles, and then learn neural models that are pose- and shape-dependent for clothing deformation. Methods like [PLPM20, WCPM18] utilize lower-dimensional representations (e.g., PCA or garment sewing parameters) to model clothing style and size, although they lack explicit control over garment fit for arbitrary body shapes. On the other hand, SIZER [TBTPM20] is the first model that learns clothing deformations explicitly as a function of clothing size, enabling controllable synthesis.

3.1.3.3 Shape under Clothing

In [BB08, ZPBPM17a, YFHWW18], the authors propose to estimate body shape under clothing by fitting a 3D body model to 3D reconstructions of people. An objective function typically forces the body to be inside clothing while being close to the skin region. These methods cannot generalize well to complex or loose clothing without additional prior or supervision [CPZ*19]. In [KBJM18, OLPM*18, XZT19, KPD19, KPBD19, XJS19], the authors propose learned models to estimate body shape from 2D images of clothed people, but shape accuracy is limited due to depth ambiguity. In this direction, our model *ParserNet* takes as input a 3D mesh and outputs 3D bodies under clothing with high fidelity while preserving subject identity (e.g., face details). We discuss this in Section 4.4. Follow-up works, such as TightCap [CPW*21] and ETCH [LFC*25], improve body shape estimation under clothing by incorporating tightness into the optimization process and utilizing a learning-based setup.

3.1.3.4 3D Cloth Parsing from Single-layered Mesh

Clothed human understanding has been widely researched in 2D and garnered success in 3D/4D, attributed to the availability of large-scale datasets [WHG*24, KKSJ24]. In particular, efficient cloth parsing in 2D has been achieved using supervised learning and generative networks [Yam12, YKB13, YL14, DLW*19, DLZ*20, GLL*18]. 3D clothing parsing of 3D scans has also been investigated [PMPHB17b, BTTPM19]. The authors propose techniques based on MRF-GrabCut [RKB04] to segment 3D clothing from 3D scans

and transfer it to different subjects. However, the approach requires several steps, which is not optimal for scalability. We introduce ParserNet in [TBTPM20], a fully automatic data-driven pipeline for parsing 3d clothing from single-layered meshes. We discuss this in Section 4.4. In [BSTPM20a], the authors jointly predict clothing and inner body surface, with semantic correspondences to SMPL. However, it lacks semantic clothing information and employs an implicit representation. Follow-up works TightCap [CPW*21] also extracts clothing from given scans. Recent learning based segmentation methods [ATO*24, MMC23, SDK*25] also parse clothing from scans, but they result in recovering partial surfaces(only the visible part). More recent works perform 4D clothing segmentation and understanding using mesh-based representation [WHG*24].

Despite significant advances in mesh-based human and clothing modeling, as well as their practical applications, they still have some limitations. Mesh-based models are limited to predefined topology for learning based tasks as discussed in [SHN*19, XYTB22]. This strictly limits the usability of such approaches to new clothing styles, especially to those with complex geometries, and does not follow the body topology. We discuss more flexible representations and their applications in human and clothing models in the next section.

3.1.4 Neural Fields for 3D Human Modeling

Neural fields are widely used for topology-agnostic and detailed 3D shape representation [PFS*19b, MON*19b, CAPM20, CMPM20], as well as for learning implicit radiance fields [MST*20] for novel view synthesis from multi-view images or videos. Both implicit surface reconstruction methods and novel view synthesis techniques have been extended to deformable objects [MON*19a, PSB*20, PCPMMN20]. These approaches introduce dynamics into neural radiance fields by encoding the scene into a canonical space using a deformation field, and then predicting density and view-dependent color values. As a result, they enable reconstruction and rendering of novel views of objects undergoing both rigid and non-rigid motions, using just a single moving camera.

In the following section, we provide an overview of neural field-based representations for modeling humans and clothing. We begin by discussing approaches that use neural fields to reconstruct human avatars from partial observations. However, these methods are often not animatable or can only be animated using an underlying body model such as SMPL. We then focus on works that incorporate body model articulation and deformation directly into the neural field formulation. We also briefly discuss modeling open surfaces such as clothing using Neural fields.

3.1.4.1 Clothed Human Reconstruction using Neural Fields

Mesh-based representations are inherently limited in terms of topology, resolution, and level of detail. This poses significant challenges for representing humans and clothing, which often involve fine-grained wrinkles, skin deformations, and dynamic movements. Early neural implicit approaches [SHN*19, SSSJ20, HZJ*21, HCJS20] aimed to reconstruct 3D clothed humans from images using neural fields. These models leverage pixel-aligned features to

improve reconstruction quality. Similarly, methods such as [CAPM20, BSTPM20a] focus on reconstructing detailed 3D geometry from partial depth observations or point clouds.

However, in both cases, the resulting models are either not animatable or are only animatable via underlying body models, such as SMPL, often using nearest-neighbor skinning weights [BSTPM20b]. This leads to limitations, especially for loose clothing or garments with large deformations, which cannot be accurately reposed using such formulations and often exhibit noticeable artifacts.

3.1.4.2 Animatable Digital Avatars using Neural Fields

Subsequent works have introduced skinning fields, pose-dependent deformations, or learned pose-independent canonical representations using neural fields to bridge the gap between standard neural representations and those better suited for articulation and deformation. The methods discussed in the following sections are either concurrent with or follow up on our contribution [TSTPM21].

Part Based Models: NASA[DLJ*20] extends the implicit surface formulation to represent articulated objects using a composition of part-based neural implicit functions. Concurrent works such as LEAP [MZBT21], NiLBS [JLJ*20], COAP [MSB*22] also use part-based neural implicit functions and rely on more sophisticated skinning and pose encoding networks. However, the part-based representation in [DLJ*20] often leads to artifacts at part boundaries and fails to capture fine details.

Learned Implicit Skinning Fields: Earlier approaches like ARCH [HXL*20] and LoopReg [BSTPM20b] project skinning weights into 3D space, but these weights are not learned—they are precomputed using nearest-neighbor strategies. In [HXL*20], clothed humans are modeled as pixel-aligned implicit functions in an unposed canonical space, but pose-dependent effects are not captured. Other works [BMTE21] learn skinning weights for diverse topologies using graph convolutional networks. SCANimate [SYMB21] also extends neural implicit models by introducing volumetric skinning weights. Yang et al. [YWM*21] propose to jointly predict shape, skeleton, and skinning weights from image or depth input to animate characters. Various formulations of volumetric skinning weights have been proposed both concurrently with and following our work [TSTPM21].

Forward Skinning and Backward Skinning: Forward skinning methods [CJS*23, CZB*21, CJS*22] define deformation in a forward manner and learn skinning weights using iterative root-finding algorithms. In contrast, backward skinning methods [TSTPM21, JLJ*20, SYMB21] learn to map points from the posed to the canonical space in a data-driven way, resulting in more detailed and dynamic reconstructions. Some recent works have also modeled dynamic deformations using autoregressive approaches [BBR*22]. We provide a comprehensive comparison and formulation of these methods in Section 5.4.

Novel View Synthesis using Animatable NeRF: Neural fields have also been extensively explored for novel view and pose synthesis tasks, particularly for human avatars. Ideas

initially developed for neural implicit surface reconstruction methods can be naturally adapted to Neural Radiance Field (NeRF)-based formulations. Recent works have utilized these concepts to achieve animatable reconstructions of human avatars from monocular or multi-view video inputs [WCS*22, ZYZ*22, IRG*23b, JYS*22]. These methods typically represent dynamic human bodies by first mapping observations to a canonical space [PCPMMN20], and then synthesizing new viewpoints or poses by applying learned deformation fields. This approach not only enables the realistic rendering of novel views but also captures intricate details and pose-dependent dynamics, providing significant advancements over traditional reconstruction methods.

3.1.4.3 Modeling Open Clothing Surfaces using Neural Fields

Although neural implicit methods have shown impressive performance in modeling high-quality, detailed surfaces, they are fundamentally limited by their reliance on watertight mesh representations. As a result, they are not well-suited for representing clothing, which often consists of open surfaces and complex topologies. To address this, several works have proposed modeling open surfaces using unsigned distance fields [CMPM20]. SMPLicit [CPA*21b] introduces a generative model for garments with varying topology using NDFs. However, extracting meshes from these representations typically requires additional post-processing, such as Poisson surface reconstruction [KBH06].

Other approaches introduce novel representations to better handle open surfaces. For instance, some works use unsigned distance fields along with their gradients [GSF22], or propose the 3-Pole SDF representation [CLLY22], both of which can be integrated with marching cubes [LC87] or its modified versions to extract surfaces. Additionally, some methods learn open surfaces using two separate fields: one to model distance, and another to explicitly capture open boundaries or holes [AWH*22, SOTC22]. Building on the meshUDF formulation [GSF22], recent works have successfully applied this approach to model clothing and other open surfaces [HYZ24, DLLG*23, LGF23, SMR*24], demonstrating its effectiveness for capturing the intricate structure and geometry of garments.

3.1.5 Point-Based Representations of Humans in 3D

Neural field-based methods, while powerful, still face significant limitations—namely, expensive inference times and limited compatibility with existing graphics pipelines and tools. These challenges motivate a renewed interest in more explicit, yet topologically agnostic, representations, as discussed in Chapter 6.

To this end, we present a comprehensive overview of point-based representations for 3D shapes. We then explore works that focus specifically on point-based models for humans and clothing. Finally, we review recent developments in 3D Gaussian Splatting (3DGS) for human avatars, highlighting how point-based representations can effectively serve as a foundation for driving such Gaussian models.

3.1.5.1 3D Point Cloud based Generative Models

Point cloud learning originates from the seminal PointNet architectures [QSMG17, QYSG17b], which introduced foundational techniques for directly processing unordered point sets. Subsequent works have proposed more sophisticated models, such as the self-attention-based Point-Transformer [ZJJ*20], which captures richer and more detailed features. KPConv [TQD*19] introduces kernel-based convolutions defined in Euclidean space, while SGPN [WYHN18] employs a single network to jointly predict point groupings and their semantic classes. Recent state-of-the-art methods [HYL*21] include DGCNN [WSL*19], which proposes EdgeConv—a differentiable layer that dynamically constructs graphs at each layer of the network. DeltaConv [WNEH22] further advances this by introducing anisotropic convolution layers. Additionally, hybrid approaches such as PVCNN [LTLH19] integrate both point-based and voxel-based representations to leverage the strengths of both paradigms.

These architectural innovations in point cloud feature learning have enabled the development of generative models for 3D point clouds. Early approaches include auto-encoding frameworks [ADMG18] and adversarial models like TreeGAN [SPK19]. More recent methods adopt a probabilistic modeling perspective. For instance, PointFlow [YHH*19] uses normalizing flows, ShapeGF [CYAE*20] leverages gradient fields, and [LH21a] applies denoising diffusion probabilistic models (DDPMs). These methods often rely on learning latent spaces through VAEs or normalizing flows and generate point clouds by modeling their distribution as a diffusion process. PVD [ZDW21] takes a different approach by jointly modeling shape latents and point distributions without separating the two. Our work, TailorDiff, builds closely on LION [ZVW*22], which utilizes diffusion models (DDPMs) within a hierarchical latent space tailored for learning generative models of clothing. We extend this framework by modifying the core PVCNN [LTLH19] architecture to explicitly incorporate body shape information into the model.

3.1.5.2 3D Point Cloud based Human Avatars

The use of point cloud generative models for representing humans and clothing is a relatively underexplored area in recent years, compared to mesh- or neural implicit-based approaches. However, a number of promising works have demonstrated the potential of point-based representations in capturing complex deformations and diverse clothing variations. Sparse, point-based representations of the human body have been leveraged in tasks such as human reconstruction from images, particularly in the context of human-object interactions. For example, InterTrack [XLPM25] and template-free methods, such as [XBLPM24], utilize point clouds to model and reconstruct interacting human subjects directly from monocular inputs without relying on predefined templates.

Animatable Digital Avatars using Point Cloud: POP [MYTB21] introduces a unified, point-based model that generalizes across multiple subjects and clothing styles. It transforms a single static 3D scan into an animatable avatar, capturing realistic pose-dependent clothing deformations. This framework highlights the expressive power of point clouds for animating clothed humans without the need for explicit mesh connectivity. Similarly, the work "Point-

Based Modeling of Human Clothing" [ZMGL21] proposes a generative model that predicts 3D point clouds representing various clothing styles conditioned on different human body shapes and poses. This enables the synthesis of diverse outfits in a topology-agnostic manner. Skirt [MYBT22] focuses on modeling garments with complex dynamics, such as skirts, using point clouds to better handle large deformations that are often difficult to capture with meshes. More recently, PocoLoco [SDL*25] presents a diffusion-based generative model that represents clothed humans as a single-layer point cloud. This method learns pose-dependent deformations in a diffusion framework, showcasing how diffusion models can effectively capture fine-grained details and motion in clothing.

Together, these works demonstrate the growing interest in and potential of point cloud-based representations for clothing and human modeling, particularly for applications requiring flexible topology, efficient learning, and high-fidelity geometric detail.

Novel View Synthesis using Animatable 3DGS: To achieve high-quality and realistic rendering of humans, NeRF [ZYZ*22] and 3DGS-based methods [KKLD23] have become popular. Animatable 3D Gaussian-based work focuses on novel views and novel pose synthesis of avatars. The primary focus of existing work is to learn animatable 3D Gaussian representations from monocular videos [LWP*23, HL23, QWM*24, HZZ*24] or multi-view videos [LZWL24, PZK*24]. These methods typically follow a common structure, where they estimate 3D Gaussian splatting parameters in canonical space, driven by SMPL [LMR*15], and enhance the output by capturing high-frequency details. This is achieved either through pose-dependent offsets [HL23, QWM*24, SMAB24], learning LBS offsets [LYXC24], or by predicting pose- and view-dependent color values [QWM*24]. GART [LWP*23] introduces an auxiliary skeleton to model complex articulations and the deformation of loose clothing. Other works rely on more predefined subject-specific templates to improve the robustness and quality of results. Both [LZWL24, PZK*24], start from a predefined template mesh for a given subject in fixed clothing and use a texel-based representation to learns 3D Gaussian parameters as an image translation task. On the one hand, this work yields very high-quality results, but it requires multiple stages of data processing, as it relies on a predefined template and takes approximately 2-4 days of training for each subject. Works like HAHA [SMAB24] and SplattingAvatar [SWL*24], rely on the more generic mesh representation(SMPL) by either using a combination of mesh+3DGaussians rasterization or by embedding gaussians on the mesh faces [SWL*24].

Novel View Synthesis using Animatable and Layered 3DGS: Although these approaches significantly improve efficiency during inference and the quality of novel view and pose synthesis, they still treat the human body and clothing as a single layer. This simplification doesn't capture the true layered nature of clothing, limiting the applicability of the learned representations to animation and novel view synthesis (NVS). Recent works, like LayGA [LLS*24] and Disco4D [PLC*24], aim to overcome these limitations by modeling humans and clothing as separate, disentangled layers. PuzzleAvatar [XYL*24] uses DMTets [SGY*21].

3.1.6 Layered and Disentangled Human in Clothing

Modeling realistic avatars requires modeling clothing along with the human body. Most prior works represent humans and clothing as a single layer [MYR*20, CJS*22, CZB*21, WCS*22, LZWL24], thereby failing to capture the layered nature of clothing and the body, which limits realism and applications such as virtual try-on and precise deformation modeling.

Disentangled Human in Clothing: Some recent works aim to disentangle humans and clothing [FLB*23, FYP*22], but still use a single combined representation for all clothing, retaining the same limitations. Other works [ZBS*25, HHH*23, DFH*24, WLD*23, LLS*24, PLC*24] represent each clothing layer separately but are designed for specific tasks. D3GA [ZBS*25] models humans and clothing with 3DGS [KKLD23] as separate layers, but cannot handle more than one clothing layer and occlusions in clothing. HumanLIFF [HHH*23], TELA [DFH*24], and SO-SMPL [WLD*23] are text-conditioned avatar creation models that generate clothing layers sequentially using diffusion. LayGA [LLS*24] uses a multi-stage process to get the 3DGS parameters of a scan and can also not model occluded clothing. These works are task-specific and cannot be applied to general clothing modeling across different modalities. GALA [KKSJ24] extracts clothing layers from scans using DMTets [SGY*21] representation, driven by the SMPL model, which enables pose editing and clothing transfer. However, this is a multi-stage process, which extracts layers one by one, and the quality of the result depends on the quality of the previous stage. Moreover, the DMTet representation is inherently unsuitable for non-water-tight surfaces like clothing. In contrast, our work TailorDiff presents a controllable and robust clothing model that can be used for applications such as clothing generation, editing, fitting to images, scans, and reconstructing clothing from occlusions *etc*.

Layered Fitting to Partial Observations: Fitting clothing or human models to scans is essential for bringing scans into correspondence for learning, and editing purposes. SMPL has been the go-to model for scan fitting [ZPBP17, MCPM24, CPMAMN22], but as noted earlier, this single-layer representation fails to capture the layered nature of clothing and the body. ClothCap [PMPHB17a], MGN [BTTPM19], SIZER [TBTPM20] use predefined clothing templates driven by SMPL to fit 3D/4D scans. However, these approaches are constrained to predefined templates and are expensive to scale to new, diverse clothing types. Clothing models such as [DLLG*23, CPA*21a, RGRF22, HYZ24] also show fitting to clothed scans, but are limited in the range of clothing styles they can represent and often lack diversity. In contrast, TailorDiff is a unified model that can represent a wide range of topologically diverse clothing within a single framework. It supports fitting to partial observations, such as scans with occluded clothing, without the need for multi-stage processing [KKSJ24].

3.2 Human Pose Priors

In this section, we present related work relevant for Chapter 7. PoseNDF is a *human pose prior* built as *neural field* in high-dimensional space. We first discuss works in human pose prior and then provide a brief discussion of follow-up works that use neural distance field-based formulation for learning the data manifold.

3.2.1 Pose and Motion Priors

Human pose and motion priors are crucial for preserving the realism of models estimated from captured data [vMHB*18a, GMSPM21, MGT*19a] and to estimate human pose from images [BKL*16, PCG*19, XZT19, CPB*20, KPBD19] and videos [KAB20, SGdA*10]. Further, they can be powerful tools for data generation. Initial work along this direction mainly focused on learning constraints for joint limits in Euler angles [ENNE12] or swing and twist representations [BB00, SNTH03, AB15], to avoid twists and bends beyond certain limits. A next iteration of methods fits a Gaussian Mixture Model (GMM) to a pose dataset and uses the GMM-based prior for downstream tasks like image-based 3D pose estimation [BKL*16, SBIK16] or registration of 3D scans [AMB*19, BTTPM19, TBTPM20]. Additionally, simple statistical models, such as PCA, have been proposed [OSBH00, UFF06, SBF00].

Task-specific Pose Priors: With the rise of deep learning and GANs [GPAM*14], adversarial training has been used to bring predicted poses close to real poses [KBJM18, GLK*20] and for motion prediction [BKL18]. However, these are task-specific models, HMR [KBJM18] models $p(\theta|I)$ and require an image I. HP-GAN [BKL18] models $p(\theta_t|\theta_{t-1})$ and requires pose parameters θ_{t-1} for previous frame/time. HuMoR [RBH*21] proposes to learn a distribution of possible pose transitions in motion sequences using a conditional VAE. It can be used for motion denoising, reconstruction from video *etc*. However, these methods are specific to a predefined set of tasks and they cannot be used as a prior for other tasks.

VAE based Pose/Motion Priors: More recent work uses VAEs to learn pose priors [PCG*19], which can be used for generating pose samples, as a prior in pose estimation, or 3d human reconstruction from images or sparse/occluded data. Some works [RBH*21, ZZB*21, PBV21] propose VAE-based human motion models. HuMoR [RBH*21] proposes to learn a distribution of possible pose transitions in motion sequences using a conditional VAE. ACTOR [PBV21] learns an action conditioned VAE-Transformer prior. Further work designs pose representations along the hierarchy of human skeletons [AKH19] and uses it for character animation [ALAC21]. All VAE-based methods rely on Gaussian assumptions in their latent space, which introduce inherent limitations – such as a bias toward sampling more likely poses and generating samples from dead regions. In contrast, our work PoseNDF [TAL*22] departs from these assumptions by directly learning the pose manifold

using neural distance fields.

GAN based Pose Prior: Concurrent work [DRC*22] learns a human pose prior using GANs and highlights the shortcomings of Gaussian assumption-based models like VPoser [PCG*19]. A VPoser decoder is used as a generator (mapping $z \rightarrow \theta$), and an HMR [KBJM18]-like discriminator is used to train the model. As mentioned, other works also rely on GAN for learning pose prior [KBJM18], but they are task-specific. Our work PoseNDF [TAL*22] follows a different paradigm than the VAE and GAN-based methods, as we directly model the manifold of plausible poses in high-dimensional space, resulting in a distance-preserving representation.

Implicit based Pose Prior: Before the rise of deep learning, modeling partial pose spaces as implicit functions was common, e.g. as fields on a single shoulder joint quaternion [HUH02] or an elbow joint quaternion, conditioned on the shoulder joint [HUF04]. However, those that ignore the real part of the quaternion, leading to ambiguities in representation, are not differentiable, and are limited to 2 joints in the human body model. In contrast, we present the first model for learning a pose prior using implicit functions. PoseNDF [TAL*22] method uses a fully differentiable neural network, which learns an implicit surface in higher dimensions, taking all human joints and all four components of each quaternion into account.

3.2.2 Neural Implicit for Learning Data Manifold

Neural fields [CMPM20, PFS*19b, MON*19b] for surface modeling have received increasing interest over the recent years. They have been used to model fields in 2D or 3D, representing images or partial differentiable equations [SMB*20, GYH*20], signed or unsigned distances from static 3D shapes [CMPM20, PFS*19b, HXL*20, HXS*21], pose-conditioned distance field [SYMB21, TSTPM21, MZBT21], radiance fields [MST*20, PCPMMN20] and more recently for human-object [XBPM22, BXP*22] and hand-object [ZBLPM22] interactions. For a more detailed overview of neural fields, please refer to [XTS*22]. Neural fields have recently been brought to higher dimensions to model surfaces in Euclidean spaces [NdSL*22]. With PoseNDF [TAL*22], we present a novel approach to learning a pose manifold using neural distance fields. In the following sections, we present subsequent works that extend neural implicit methods for learning data manifolds.

Motion and Pose Priors using NDFs: In recent times, several works have explored learning data manifolds using neural distance fields or similar implicit representations. NRDF [HTB*24] learns implicit manifolds of plausible articulated poses by leveraging geodesic distances in quaternion space. NRDF is built on top of PoseNDF [TAL*22], a method presented in Chapter 7. This work also draws connections between NRDF and their resemblance to Riemannian Flow Matching (RFM) [CL24], an extension of the state-of-the-art generative model Flow Matching. The key differences lie in how training data and time steps are sampled, as well as in the computation of the flow. Momanifold [DFZ*24]

introduces a neural implicit approach that decouples human motion dynamics into joint-wise acceleration manifolds. This method explicitly quantifies motion plausibility and efficiently handles high-dimensional motion data, outperforming previous motion priors across various downstream tasks. Similarly, [GNF*25] employs neural distance fields with a polar coordinate-based representation to model 2D human pose priors. It introduces a specialized non-geodesic distance metric and a gradient-based data augmentation strategy, significantly improving pose estimation refinement.

Generic Data Priors using NDFs: All the aforementioned works are still limited to human pose and motion data. A particularly interesting direction is taken by EBIM [RLGCC23], which proposes Energy-Based Implicit Manifolds to model complex data topologies as zero-level sets of neural networks. EBIM uses constrained Langevin dynamics for density estimation and is capable of learning intricate manifold-supported distributions more effectively than traditional pushforward methods. This work highlights the broader applicability and potential of distance-based formulations for learning generic data manifolds.

I
Human in Clothing

Realistic representation of digital humans requires accurately modeling both human bodies and their clothing, particularly given the intricate and subtle nature of clothing deformation as it interacts with varying body shapes, sizes, and motions. An ideal digital human and clothing model should be robust, user-friendly, computationally efficient, and capable of capturing the immense diversity and scale present in real world. This research focuses on developing data-driven human-in-clothing models by learning from extensive real-world and large-scale synthetic datasets, with particular emphasis on representations that are conducive to learning-based approaches.

In **Chapter 4**, we present SizerNet, the first size-sensitive 3D clothing model based on mesh representation. Our focus is on the challenging task of learning subtle and non-linear clothing deformations that occur with changes in clothing size for a given garment. As a significant contribution to the research community, we introduce the first large-scale dataset of 3D humans in clothing, capturing variations in both clothing size and body shape. We also propose ParserNet, the first learning-based method for parsing clothing layers from single meshes, enabling clothing editing using our SizerNet model.

In **Chapter 5**, we address the limitations of template-based methods – such as intensive data processing, limited fidelity – and turn to neural field-based representations. We propose NeuralGIF, a novel neural field formulation designed specifically to model detailed, pose-dependent clothing and skin deformations. Our approach introduces implicit skinning fields, enabling the transition from discrete to continuous, volumetric human body representations that can model not only unclothed human bodies but also complex, clothed humans.

In **Chapter 6**, we move towards a unified clothing model capable of representing all topologically varying clothing within a single framework – not addressed in prior work. We propose TailorDiff, the first large-scale, unified, and generative clothing model that is gender-agnostic, compatible with body models, and based on a point cloud representation. Inspired by the simplicity and utility of the SMPL body model, our approach serves as a SMPL-like foundation model for clothing. TailorDiff is explicit, seamlessly integrates with standard graphics pipelines, supports multiple clothing categories, enables layered clothing fitting, and is fully differentiable – making it useful for real-world digital human applications.

4

SIZER: Parsing and Learning Size Sensitive 3D Clothing



Fig. 4.1 **SIZER**: A dataset and model of people with clothing size variation. *Left*: We present the first dataset of 3D scans of people captured in different clothing **styles and sizes**. *Right*: We further introduce clothing parsing and resizing in 3D. Specifically, given a registered scan of a subject, **ParserNet** can segment the upper and lower clothing in 3D, and **SizerNet** can resize the parsed clothing to a desired size. Here, we present the ground truth 3D scan with texture, along with the parsed and resized clothing behind the scan, which illustrates the subtle differences introduced by changes in clothing size when draped on the same body.

Realistic deformation of clothing has been crucial for providing a more immersive and smooth experience for virtual avatars in AR/VR, content generation, and virtual try-on applications. To learn the intricacies of clothing deformation, researchers have often relied on data-driven methods learned from real-world data. A significant amount of work has been dedicated to learning body shape- and body pose-dependent or clothing material-dependent deformation. However, a very crucial aspect of variation has not been addressed sufficiently – clothing size as we observe it in the real world. Changes in the shape of clothing with respect to changes in size labels are subtle and often depend on the style of the clothing and the individual's body shape. Moreover, most existing works either treat the human and clothing together as a single connected surface in 3D or focus on individual clothing items designed synthetically. In this chapter, we address both of these challenges. First, given a scan of a clothed person represented as a single mesh (with fixed topology), we parse each clothing layer (namely, upper and lower garments) and the underlying body in 3D, followed by learning a size-dependent clothing deformation model.

Specifically, we presented SizerNet, a model that predicts 3D clothing based on human body shape and garment size parameters, and ParserNet, a model that infers garment meshes and the body shape under clothing, including personal details, in a single pass from an input mesh. SizerNet enables the estimation and visualization of the dressing effect of garments across various sizes, while ParserNet facilitates direct editing of clothing on an input mesh, eliminating the need for scan segmentation —a challenging problem in itself.

To learn these models, we introduced the SIZER dataset, which includes approximately 2,000 scans of 100 different subjects wearing casual clothing items in various sizes. The dataset provides raw scans, registrations to the SMPL model, scans segmented into clothing parts, and labels indicating garment category and size.

Our experiments demonstrated that SizerNet and ParserNet achieved better parsing accuracy and more reliable size prediction compared to baseline methods trained on the SIZER dataset. The code, models, and dataset were made publicly available for research purposes.

This chapter is based on **SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing, Tiwari *et al.*, ECCV202** [TBTPM20]. This work was accepted as an Oral presentation(Top 2%). It provides the first large-scale dataset of diverse people in diverse clothing, and has since been widely used for benchmarking tasks such as 3D reconstruction of humans from images [XXMPM24, XYTB22, XYC*23], SMPL registration [CPMAMN22], clothing segmentation [MMC22, MMC23, ATO*24], and layered clothing reconstruction from scans or images [DLLG*23, CPA*21a]. With this work, we introduce the first size-sensitive clothing model in 3D, which can model subtle variations *wrt* clothing sizes and body shapes.

4.1 Introduction

Modeling how 3D clothing fits on the human body as a function of size has numerous applications in 3D content generation (e.g., AR/VR, movies, video games, sports), clothing size recommendation (e.g., e-commerce), computer vision for fashion, and virtual try-on. It is estimated that retailers lose up to $600 billion each year due to sales returns, as it is difficult to purchase clothing online without knowing how it will fit [ihl, retb]. The subtlety in changes in clothing deformation or fitting as the size changes makes this problem even more challenging. AI-driven solutions have helped [reta] reduce this impact by providing virtual try-on solutions, better size recommendations, and better trend recommendations. More work has been done in this domain from research [CCSC23, YNU24] and industrial [Tru25, 3DL25] perspectives.

Predicting how clothing fits as a function of body shape and clothing size was an extremely challenging task. Clothing interacts with the body in complex ways, and fit is a non-linear function of size and body shape. Furthermore, *clothing fit differences with size were subtle*, but they could make a difference when purchasing clothing online. Physics-based simulation has been the most commonly used technique because it generalizes well, but unfortunately, it is difficult to adjust its parameters to achieve a realistic result, and it can be computationally expensive and requires expertise.

While several works in the past have focused on how clothing deforms as a function of pose [LCT18a], or pose and shape [LCT18a, SOC19, GCS*19, PLPM20, MYR*20], size-dependent clothing deformation remained underexplored and a challenging task. Post-2020, following the publication of this work [TBTPM20], size-based virtual try-on was increasingly explored in the image domain [CCSC23, YNU24, KNF23]. In this setting, given an image of a person wearing clothing, methods can generate how the clothing would appear in different sizes. However, most of these approaches focused primarily on upper-body clothing, such as shirts and t-shirts. Furthermore, since these methods operated directly on images, they were inherently limited — they could not easily model how resized clothing would behave under new poses or motion. Synthetic data-driven works learn a space of styles [WCPM18, PLPM20] from physics simulations, but they aim at plausibility, and therefore cannot predict how *real clothing* will deform on a real body under subtle size variations. Hence, developing a model that can resize clothing in 3D, across a wide range of clothing types, while producing a physically plausible, simulation-ready representation, is crucial for advancing virtual try-on technologies. Such a model would not only enable realistic visualization of size variations across different body shapes but also allow for dynamic interactions, motion synthesis, and seamless integration into AR/VR platforms, gaming, and e-commerce applications.

Precisely predicting how clothing deforms with subtle size variations requires capturing interactions between the clothing layer and the underlying body, something previous approaches lacked due to limited real-world datasets and insufficient modeling techniques. To address the critical challenge of modeling how clothing fits across varying body shapes and clothing sizes, we introduced (1) a 3D dataset of people wearing the same clothing in different sizes and (2) a data-driven model *learned from real scans* which varies with sizing and body shape.

Firstly, we introduced the *SIZER* dataset, the first extensive dataset specifically designed for modeling clothing deformation as a function of size. It contains approximately 2000 high-quality 3D scans featuring 100 diverse subjects, each wearing 10 different clothing in four standardized sizes (S, M, L, XL). We provided detailed annotations, including 3D clothing segmentations, SMPL+G [BTTPM19, PMPHB17b] registrations, estimates of the underlying body shapes, and explicit clothing class and size labels. This dataset not only bridged the gap in available data for clothing size modeling but also facilitated research and benchmarking for 3D human reconstruction [XXMPM24], 3D clothing modeling [CPA*21a, DLLG*23, LGF23], and 3D clothing segmentation [MMC22, MMC23, ATO*24, SDK*25].

Secondly, leveraging this unique dataset, we propose *SizerNet*, the first neural network model trained on real-world scans to predict clothing deformation as a function of both body shape and clothing size. SizerNet employs an encoder-decoder architecture, where the encoder maps an input multi-layered mesh into a compact latent representation, and the decoder synthesizes realistic clothing shapes conditioned on SMPL body parameters, the clothing's original size, and the desired output size. Unlike previous approaches relying solely on synthetic data, our model generalizes effectively to real clothing and accurately captures subtle size-dependent fit variations, thus significantly enhancing realism in virtual try-on applications. Furthermore, since SizerNet models resize using a mesh-based representation,

it can be seamlessly integrated with state-of-the-art clothing simulation methods to visualize the dynamic behavior of the resized 3D clothing on various poses—an essential aspect for assessing clothing fit on the human body.

Real-world 3D data, such as scans captured by specialized scanners or reconstructed from images and videos, typically comes as single-layered mesh representations [SHN*19]. However, in order to do the resizing of clothing and editing tasks, it is crucial to disentangle the clothing layers from the underlying body—a process known as *3D clothing parsing*. To address this challenge, we introduced *ParserNet*, a fast feed-forward framework which automatically maps a single mesh registration (SMPL deformed to the scan) to semantically meaningful multi-layer meshes, namely body, upper clothing, and lower clothing. ParserNet efficiently transforms these single-layer meshes (SMPL body meshes aligned to the scans) into structured, multi-layered mesh representations, clearly separating the body shape from clothing layers. *ParserNet*, not only segments the single mesh registration, but it reparameterizes the surface so that it is coherent with common clothing templates. In contrast to prior methods [BTTPM19, PMPHB17a] that rely heavily on computationally intensive segmentation or require manual corrections, ParserNet achieves coherent segmentation and reparameterization aligned with standard clothing templates via a single feed-forward pass. The resulting multi-layer representation significantly enhances flexibility, enabling seamless clothing resizing, direct editing, and dynamic simulation, thus broadening the potential applications in virtual try-on, e-commerce, gaming, and AR/VR. The challenge of disentangling multi-layered representation from easily available data is still challenging and has not been fully addressed. We discuss more about this in Section 4.7 and further in Chapter 6.

Additionally, the tandem of *SizerNet* and *ParserNet* allows us to edit the size of clothing directly on the mesh, allowing shape manipulation applications never explored before. We show the pipeline of the tandem of *SizerNet* and *ParserNet* in Figure 4.2 and show results for diverse clothing parsing and resizing in Section 4.6.2.1.

In summary, our contributions are:

- **SIZER** dataset: A dataset of clothing size variation of approximately 2000 scans, including 100 subjects wearing 10 clothing classes in different sizes, where we make available scans, clothing segmentation, SMPL+G registrations, body shape under clothing, clothing class, and size labels.

- **SizerNet**: The first model learned from real scans to predict how clothing drapes on the body as a function of size.

- **ParserNet**: A data-driven model to map a single mesh registration into a multi-layered representation of clothing without the need for segmentation or non-linear optimization.

In this chapter, we first discuss the 3D representation used in this work, which is based on the SMPL topology. We explain how SMPL can be extended to represent clothing as a separate layer in Section 4.2. In Section 4.3, we describe the *SIZER* dataset, which was captured to study clothing resizing in real-world scenarios. We then present the two models proposed in this work. Specifically, in Section 4.4, we introduce ParserNet, which learns

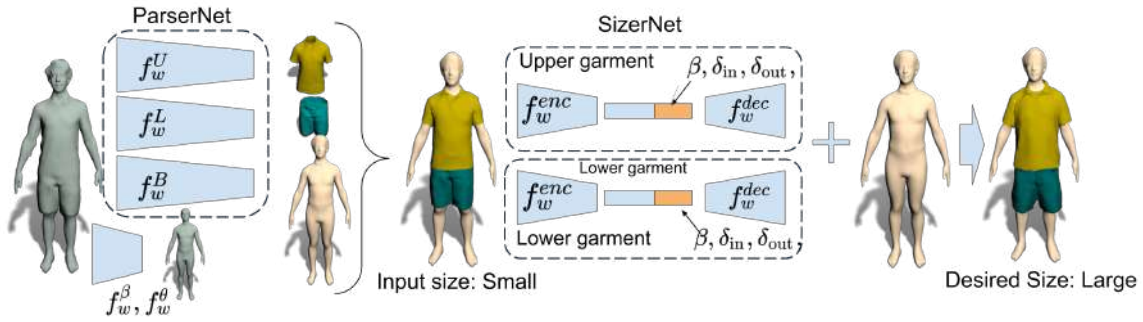


Fig. 4.2 **Clothing Parsing and Resizing Model:** We propose a model to estimate and visualize the dressing effect of clothing conditioned on body shape and clothing size parameters. For this we introduce *ParserNet* (f_w^U, f_w^L, f_w^B), which takes a SMPL registered mesh $M(\theta, \beta, \mathbf{D})$ as input and predicts the SMPL parameters (θ, β), parsed 3D clothing using predefined templates $T^g(\beta, \theta, \mathbf{0})$ and predicts body shape under clothing while preserving the personal details of the subject. We also propose *SizerNet*, an encoder-decoder ($f_w^{\text{enc}}, f_w^{\text{dec}}$) based network, that resizes the clothing given as input with the desired size label $(\delta_{\text{in}}, \delta_{\text{out}})$ and drapes it on the body shape under clothing.

to extract a multi-layered representation of clothing and the body underneath from single-layered surfaces. In Section 4.5, we detail the clothing resizing model we propose. This is followed by an experimental analysis of both proposed methods in Section 4.6. Finally, in Section 4.7, we discuss the limitations of our approach and the chosen representation. We also reflect on the impact of our model and dataset in follow-up works and motivate the exploration of alternative representations for modeling *Humans in Clothing*, which we investigate in the following chapters.

4.2 Mesh-based Clothed Human Representation

Accurate and expressive representation of clothed humans from real-world data is a challenging task due to the diversity present in clothed humans. As introduced in Chapter 1, throughout this thesis, we systematically explore and analyze different representations for clothed human modeling with respect to key criteria such as scalability, versatility, representational power, ease of learning, and applicability. In this particular work, we take a deep dive into the mesh-based representation. Specifically, meshes with fixed, coherent topology have been the standard representation due to their widespread applicability in animation, rendering, and learning-based methods.

Among mesh-based models, the SMPL [LMR*15, PCG*19] body model provides a strong foundation owing to its parametric formulation, consistent vertex ordering, and support for articulated pose and shape deformations. However, vanilla SMPL lacks the ability to express personalized surface-level variations such as clothing geometry or hairstyles. To address this limitation, we explore enhanced formulations that extend SMPL with additional deformation components to better capture such details.

We first present the SMPL+D representation in Section 4.2.1, which augments the base mesh with per-vertex displacements to encode clothing and identity-specific features. Following this, we describe the SMPL+G representation in Section 4.2.2, which introduces explicit submeshes for different clothing categories, allowing for semantic disentanglement of clothing from the underlying body.

4.2.1 Clothed Human Representation via Vertex Displacements

To represent clothing and personal features such as hairstyles and body-specific details, we use SMPL+D representation. SMPL+D representation augments the SMPL body model by introducing per-vertex displacements (**D**) on top of the SMPL mesh. The SMPL body model itself provides a parametric formulation defined as a function $M(\cdot)$ of pose (θ) and shape (β), leveraging benefits such as coherent topology, pose articulation, and compatibility across various subjects. Formally, the displacement-augmented model, referred to as SMPL+D, extends the original SMPL formulation as:

$$M(\beta,\theta,\mathbf{D}) = W(T(\beta,\theta,\mathbf{D}),J(\beta),\theta,\mathbf{W}), \tag{4.1}$$

$$T(\beta,\theta,\mathbf{D}) = \mathbf{T} + B_s(\beta) + B_p(\theta) + \mathbf{D}. \tag{4.2}$$

Similar to the SMPL body model, SMPL+D applies standard skinning $W(\cdot)$ to a base template **T** in T-pose. **W** denotes the blend weights and $B_p(\cdot)$ and $B_s(\cdot)$ models pose and shape dependent deformations respectively in Equation (4.2) and Equation (4.1). For more details, refer to Section 2.1.2. In this approach, all clothing and personal-specific features are encoded implicitly as deviations from the canonical SMPL template. This representation, however, lacks semantic understanding since clothing and body features are merged into a single mesh without explicit separation into distinct semantic layers.

4.2.2 Clothing Representation via SMPL+G Submeshes

To address the absence of semantic and layered clothing information in SMPL+D, we propose to use SMPL+G [BTTPM19, PMPHB17a], a parametric model explicitly designed to represent clothing as separate semantic layers, each defined as a submesh of the SMPL model. This formulation enables explicit representation, manipulation, and understanding of clothing independently from the underlying body.

Given a scan of a clothed human, we first segment the scan into clothing and skin regions using an improved segmentation pipeline based on the Human Parsing network [GLL*18], refined by employing a multi-view consistency loss. The segmented clothing is then registered to clothing-specific templates using a multi-mesh registration approach as introduced in [BTTPM19]. The SMPL+G representation inherently supports only clothing that follows the SMPL topology, and MGN [BTTPM19] is limited to registering a fixed set of clothing classes, such as *T-shirts, shirts, full pants, short pants, and coats with shirts*. In this work, we extend the *multi-mesh registration* approach to also include *skirts* in the list of supported clothing classes. Each clothing-specific template mesh is a subset of the SMPL template mesh, defined by:

$$T^g(\beta,\theta,\mathbf{0}) = \mathbf{I}^g T(\beta,\theta,\mathbf{0}), \tag{4.3}$$

where $\mathbf{I}^g \in \mathbb{Z}_2^{m_g \times n}$ is a binary indicator matrix identifying clothing vertices within the SMPL mesh. Here, $\mathbf{I}^g_{i,j} = 1$ indicates that clothing vertex $i \in 1 \dots m_g$ corresponds to the body vertex $j \in 1 \dots n$.

The clothing mesh for each class g is parametrically defined by vertex offsets ($\mathbf{D}^g$) from its clothing-specific template:

$$G(\beta,\theta,\mathbf{D}^g) = W(T^g(\beta,\theta,\mathbf{D}^g), J(\beta), \theta, \mathbf{W}). \tag{4.4}$$

This approach maintains the advantages of SMPL, such as shared topology, coherent pose deformation, and parametric body shape representation, while allowing clothing to be individually analyzed, synthesized, and manipulated as semantic entities. For each clothing class, $\mathbf{I}^g$ is defined manually and hence suffers from scaling to novel clothing classes.

4.3 SIZER Dataset

In this work, we tackled the challenging problem of modeling clothing fitting as a function of body shape and clothing size. As previously discussed, a key obstacle in this domain is the lack of real-world datasets that capture the same clothing in different sizes, worn by real humans with diverse body shapes. To address this gap, we introduce the SIZER dataset: a large-scale, calibrated, and well-annotated 3D scan dataset of clothed humans in varying clothing sizes.

SIZER contains over **2000 high-resolution 3D scans of 100 subjects, each wearing 10 distinct clothing types in up to four sizes (S, M, L, XL)**. It is the first real-world dataset that systematically studies the effect of clothing size on fit and appearance, while controlling for body shape and pose. In Section 4.3.1, we discuss scan capture setup and registrations to SMPL and SMPL+G representations, enabling learning-based applications, followed by a detailed analysis of clothing sizing in Section 4.3.2.

4.3.1 Dataset Overview

The SIZER dataset is a large-scale collection comprising over 2,000 high-resolution 3D scans of humans wearing clothing in different sizes. It consists of 100 unique subjects, each captured in multiple clothing classes and 2–3 clothing sizes (e.g., S, M, L, XL). We focus on 10 commonly worn clothing types—t-shirt, polo shirt, dress shirt, hoodie, coat, vest, skirt, jeans, shorts, and shirt—chosen to reflect everyday fashion diversity. Figure 4.1 shows some examples of people wearing a fixed set of clothing in different sizes. To control for pose-induced artifacts, all subjects are scanned in a canonical A-pose, which minimizes stretching or tension in the clothing and ensures consistent shape comparison across sizes.

Capture Setup: We capture the dataset using Treedy's full-body static scanner, which consists of over 130 calibrated RGB cameras positioned around the subject. The scans



Fig. 4.3 **Dataset capture setup:** Sizer dataset scans were captured using Treddy static scanner [tre] with 130 fixed cameras. We show some random camera frames from the capture setup.

are reconstructed using Agisoft Metashape [agi], producing detailed mesh representations of the clothed subjects. These meshes vary in topology and resolution, posing challenges for direct use in learning frameworks. We preprocessed our dataset by registering it to SMPL [LMR*15]. We build upon the dataset captured in [Tiw20] and increase the scale and quality of the dataset. For more details, refer to [Tiw20].

To improve general usability of the *SIZER* dataset, we provide SMPL+G registrations [LIPM19a, BTTPM19] registrations(as mentioned in Section 4.2.2). Registering our scans to SMPL brings all our scans into correspondence and provides more control over the data via pose and shape parameters from the underlying SMPL. For each scan, we also provide clothing segmentation labels into three classes: upper clothing, lower clothing, and body. This is similar to clothing segmentation introduced in [BTTPM19]. We briefly describe the SMPL, SMPL+D, and SMPL+G formulations in Section 4.2. For more details on the registration pipeline, please refer to [BTTPM19, PMPHB17a].

To facilitate broader usage, we released SMPL and SMPL+G registrations [LIPM19a, BTTPM19], clothing class names, and sizes for every scan. This brings all scans into correspondence and allows parametric control over pose and shape. Each sample in the *SIZER* dataset includes:

- High-resolution 3D scan mesh with texture maps.
- Segmented clothing mesh.
- SMPL, SMPL+D, SMPL+G registration at 2 resolution levels.
- Clothing category label.
- Clothing size label.

Dataset	Static/4D	Clothing types	Clothing Sizes	# Subjects	Pose Var.	SMPL Registration
People Snapshot [AMX*18b]	3D	2-4	✗	24	✓	SMPL(+D)
Dyna [PMRMB15]	4D	✗	✗	10	✓	SMPL
BUFF [ZPBP17]	Static	2	✗	6	✓	SMPL
MGN [BTTPM19]	Static	6	✗	96	✗	SMPL(+D,G)
CAPE [MYR*20]	4D	4	✗	15	✓	SMPL
SIZER (Ours) [TBTPM20]	Static	**10**	✓	**100**	✗	SMPL(+D,G)

Table 4.1 **Comparison of SIZER with existing datasets:** We compare with 3D and 4D human/clothing datasets published prior to its release. SIZER dataset uniquely combines a wide range of clothing types with size variation across a large number of subjects, enabling size-sensitive clothing modeling.

This dataset can be used in several applications like virtual try-on, character animation, learning generative models, data-driven body shape under clothing, size, and shape-sensitive clothing model, etc. We provide a comparison of the SIZER dataset with datasets released prior to it in Table 4.1. Clothing type in this table roughly refers to a coarse style of clothing e.g., t-shirts, with slightly different geometry and new textures are referred to as one clothing type.

Following its release, the SIZER dataset has established itself as a benchmark in the field, driving progress in multiple domains, including 3D human reconstruction from images [XXMPM24, CPA*21a], parametric clothing fitting to scans or images [DLLG*23], and 3D clothing segmentation [MMC22, ATO*24]. Its scale, diversity, and high-quality annotations have made it a go-to resource for evaluating and developing models related to clothing.

4.3.2 Clothing Size Variation in Real World

As mentioned at the beginning of this chapter, the geometric variation across clothing sizes, though driven by physical growth and design intent, is often subtle and non-linear. As shown in Table 4.2 and Table 4.3, we quantify these changes both in terms of per-vertex displacements obtained via SMPL-based registration and relative changes in surface area across standard clothing sizes. These statistics are computed on high-quality ground-truth 3D scans and their SMPL+G registrations across different clothing classes.

In garment sizing, key geometric trends reveal that changes across sizes are non-linear, particularly in how surface area and vertex displacement evolve. When skipping sizes—such as jumping from size S to L instead of going stepwise through M—the increase in surface area approximately sums up the individual increments (e.g., S $\rightarrow$ M and M $\rightarrow$ L). However, vertex displacement does not follow the same additive pattern. For instance, a polo T-shirt shows a 17.8% area increase and 29.1 mm vertex displacement from S to M, and 16.0% and 29.6 mm from M to L. But jumping directly from S to L results in a 35.3% area increase (close to the sum), while the displacement is only 43.5 mm—much less than the sum of 58.7 mm. This indicates a saturation effect in geometry: once an initial shape expansion occurs, further size increases cause diminishing additional movement at each point.

The overall shape or style of the clothing also plays a crucial role. Loose garments such as shorts, hoodies, and coats scale more aggressively than fitted ones like pants or vests. For example, shorts show a significant vertex displacement of 44.9 mm with a 20.6% area change from S to M, whereas pants only register 29.1 mm and 12.8%. This is because loose clothing is designed to follow the looseness pattern with an increase in size and has more surface extending away from the body; therefore, minor circumference changes result in larger vertex shifts. In contrast, tight-fitting garments stay close to the body, limiting spatial movement with size increase.

Interestingly, surface area increase and vertex displacement are not proportional to each other. Coats expanding from M to XL exhibit an average vertex shift of 53 mm but only a 24% surface area growth. Shorts, by contrast, register a 46 mm displacement alongside a 39% area increase. This disparity reflects the fact that while surface area captures overall fabric growth, vertex displacement is more sensitive to the garment's extremities and complex shaping, with the ratio influenced by the geometry's intricacies.

Finally, early size increments often introduce sharper changes, particularly in fitted garments where the percentage area growth is steeper from S to M and flattens in larger sizes. However, this trend is not universal. Some loose garments like Shorts 2 show dramatic jumps at higher sizes—for instance, increasing 19.4% from M to L but surging 38.1% from M to XL. This reinforces the idea that garment geometry evolves in varied, class-dependent ways rather than following a consistent linear path.

These trends collectively highlight two critical challenges for data-driven modeling of clothing size variation: the subtlety and sparsity of geometric changes, especially when limited to class-specific clothing or slim-fit types, and diverse deformation characteristics across classes, making it difficult for a single model to uniformly capture size-induced deformations across clothing types. Hence, while training deformation fields or latent space models to represent size changes, it becomes essential to either condition explicitly on clothing class or build modular architectures that can adapt to the structural properties of each category. In this work [TBTPM20], we propose a modular structure for different types of clothing items.

4.4 3D Human Parsing

We previously discussed that real-world 3D scan data often comes as a single-layered mesh representation of a human in clothing. However, in reality, the clothing and human body are separate surfaces. To learn the behavior of clothing independently and to enable editing and retargeting applications, a fast method for disentangling the 3D layered representation of clothing and body is required. We introduced *ParserNet*, the first method for extracting clothing directly from SMPL-registered meshes. We first explain the key idea of *Parsing 3D Clothing* in Section 4.4.1. Then we discuss a pose and shape estimation network along with *Body ParserNet*, which is used to get detailed *body shape under clothing* in Section 4.4.2.

Clothing Item	**From Size S**		**From Size M**		**From Size L**
	S → M	S → L	M → L	M → XL	L → XL
Polo T-Shirt	29.07	43.49	29.61	36.23	28.19
T-Shirt	28.19	44.35	27.53	35.21	26.37
Shirt	32.66	44.27	32.23	43.24	30.59
Coat	32.52	47.38	30.43	53.17	33.66
Hoodies	36.02	49.22	34.67	47.35	33.21
Vest	25.20	42.62	26.50	53.42	29.72
Pants	29.08	35.33	28.89	35.87	25.52
Shorts	44.94	51.29	34.80	45.82	39.15
Shorts2	18.07	35.32	19.53	34.47	29.40

Table 4.2 **Clothing size variation using per-vertex differences:** Average per-vertex displacement (in mm) between clothing items of different sizes. The differences are nonlinear and vary with both the source size and the type of clothing. Transitions from smaller sizes to larger ones, especially for loose clothing like shorts and hoodies, show larger vertex displacements than transitions between adjacent larger sizes.

4.4.1 3D Clothing Parsing

Parsing clothing or body with personal features such as faces/hairs from a single mesh ($\mathbf{M}$) can be done by segmenting it into separate clothing for each class ($\mathbf{G}^{g,k}_{\text{seg}}$), which leads to different underlying graph connectivity ($\mathcal{G}^{g,k}_{\text{seg}} = (\mathbf{G}^{g,k}_{\text{seg}}, \mathbf{E}^{g,k}_{\text{seg}})$) across all the instances (k) of a clothing class g, shown in figure 4.4 (right). Moreover, using segmentation, we only get partial visible surfaces. Hence, we propose to parse clothing by deforming vertices of a pre-defined template $T^g(\beta, \theta, \mathbf{0})$ with fixed connectivity $\mathbf{E}^g$, obtaining vertices $\mathbf{G}^{g,k} \in \mathcal{G}^{g,k}$, where $\mathcal{G}^{g,k} = (\mathbf{G}^{g,k}, \mathbf{E}^g)$, shown in figure 4.4 (middle).

Our key idea is to predict the deformed vertices $\mathbf{G}^g$ directly as a convex combination of vertices of the input mesh $\mathbf{M} = M(\beta, \theta, \mathbf{D})$ with a learned sparse regressor matrix $\mathbf{W}^g$, such that $\mathbf{G}^g = \mathbf{W}^g\mathbf{M}$. Specifically, *ParserNet* predicts the sparse matrix ($\mathbf{W}^g$) as a function of input mesh features (vertices and normals) and a predefined per-vertex neighborhood ($\mathcal{N}_i$) for every vertex i of clothing class g. We will henceforth drop $(.)^{g,k}$ unless required. In this way, the output vertices $\mathbf{G}_i \in \mathbb{R}^3$, where $i \in \{1, \ldots, m_g\}$, are obtained as a convex combination of input mesh vertices $\mathbf{M}_j \in \mathbb{R}^3$ in a predefined neighborhood ($\mathcal{N}_i$).

$$\mathbf{G}_i = \sum_{j \in \mathcal{N}_i} \mathbf{W}_{ij}\mathbf{M}_j. \tag{4.5}$$

Rather than regressing a dense displacement field via a fully connected MLP, our method predicts a sparse matrix that linearly combines input mesh vertices to reconstruct each clothing vertex. This formulation enhances both locality and interpretability: each row of the sparse matrix clearly delineates which input vertices contribute to a given clothing vertex, effectively replacing the black-box nature of traditional networks with an explicit,

Clothing Item	Source Size S		Source Size M		Source Size L
	S → M	S → L	M → L	M → XL	L → XL
Polo T-Shirt	17.83	35.34	16.06	33.11	15.38
T-Shirt	18.53	34.54	13.56	27.46	14.21
Shirt	16.44	29.69	11.37	24.26	15.42
Coat	12.68	26.33	14.33	24.46	13.67
Hoodies	14.57	23.95	8.82	23.46	12.45
Vest	14.02	21.62	7.31	21.09	12.85
Pants	12.83	18.56	11.18	21.87	11.26
Shorts	20.58	40.46	17.99	39.23	16.79
Shorts2	13.35	34.03	19.42	38.07	20.07

Table 4.3 **Clothing size variation using surface area:** Percentage change in clothing surface area across different size transitions. The change is nonlinear and varies significantly depending on both the source and target sizes as well as the clothing type.

geometry-aware mapping. Additionally, by reusing geometric information directly from the scan instead of generating new geometry, our approach inherently preserves fine details.

4.4.2 Body Shape under Clothing

To estimate body shape under clothing, we first create the undressed SMPL body for a given clothed input single-layer mesh $M(\beta,\theta,\mathbf{D})$, by predicting θ,β using f_w^θ and f_w^β respectively. We train f_w^θ and f_w^β with L_2 loss over parameters and per-vertex loss between predicted SMPL body and clothed input mesh, as shown in equations (4.6) and (4.7). Since the reference body under clothing parameters θ,β obtained via instance specific optimization (Section 4.3.1) can be inaccurate, we add an additional per-vertex loss between our predicted SMPL body vertices $M(\hat{\theta},\hat{\beta},\mathbf{0})$ and the input clothed mesh $M(\beta,\theta,\mathbf{D})$. This brings the predicted undressed body closer to the input clothed mesh. We observe more stable results training f_w^θ and f_w^β separately initially, using the reference β and θ respectively. Since the β components in SMPL are normalized to have $\sigma = 1$, we un-normalize them by scaling by their respective standard deviations $[\sigma_1,\sigma_2,\ldots,\sigma_{10}]$ as given in Equation (4.7).

$$\mathcal{L}_\theta = w_{\text{pose}}||\hat{\theta}-\theta||_2^2 + w_v||M(\beta,\hat{\theta},\mathbf{0}) - M(\beta,\theta,\mathbf{D})|| \tag{4.6}$$

$$\mathcal{L}_\beta = w_{\text{shape}}\sum_{i=1}^{10}\sigma_i(\hat{\beta}_i-\beta_i)^2 + w_v||M(\hat{\beta},\theta,\mathbf{0}) - M(\beta,\theta,\mathbf{D})|| \tag{4.7}$$

Here, w_{pose}, w_{shape} and w_v are weights for the loss on pose, shape, and predicted SMPL surface. $(\hat{\theta},\hat{\beta})$ denote predicted parameters. The output is a *smooth* (SMPL model) body shape under clothing.

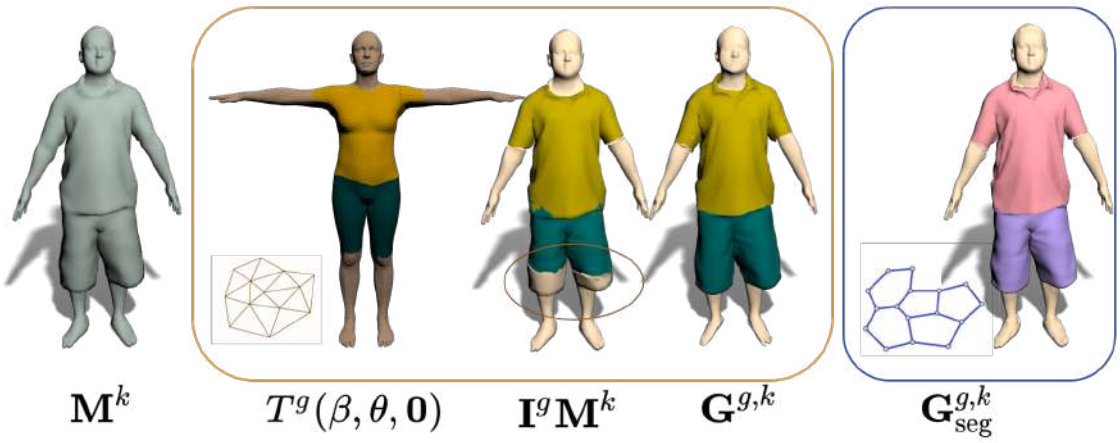


Fig. 4.4 **3D Clothing Parsing from single mesh:**(*Left to right*) Input single mesh ($\mathbf{M}^k$), clothing template ($T^g(\beta,\theta,\mathbf{0}) = \mathbf{I}^g T(\beta,\theta,\mathbf{0})$), clothing mesh extracted using $\mathbf{G}^{g,k} = \mathbf{I}^g\mathbf{M}^k$, multi-layer meshes ($\mathbf{G}^{g,k}$) registered to SMPL+G, all with clothing class specific edge connectivity $\mathbf{E}^g$, and segmented scan $\mathbf{G}^{g,k}_{\text{seg}}$ with instance specific edge connectivity $\mathbf{E}^{g,k}_{\text{seg}}$.

Parsing detailed body shape under clothing: For generating detailed body shape under clothing, we first create a *smooth body mesh*, using SMPL parameters θ and β predicted from $f^{\theta}_{w}, f^{\beta}_{w}$ (Section 4.4.2). The predicted body shape and pose parameters give a good estimate of the body underneath. However, this doesn't include personalized subject features such as facial features or hair *etc*. We want to use the aforementioned convex combination formulation. *Body ParserNet* transfers the visible skin vertices from the input mesh to the smooth body mesh, thereby obtaining hair and facial features. We parse the input mesh into upper, lower clothing, and detailed shape under clothing using 3 sub-networks ($f^{U}_{w}, f^{L}_{w}, f^{B}_{w}$) of *ParserNet*, as shown in Figure 4.2.

4.4.3 Training and Implementation Details

Loss Functions: We train *ParserNet* with training loss given by Equation (4.8), where $w_{3D}, w_{\text{norm}}, w_{\text{lap}}, w_{\text{interp}}$ and w_{w} are weights for the loss on vertices, normal, Laplacian, interpenetration and weight regularizer term respectively. We explain each of the loss terms in this section.

$$\begin{aligned}\mathcal{L}_{\text{parser}} =& w_{3D}\mathcal{L}_{3D} + w_{\text{norm}}\mathcal{L}_{\text{norm}} + w_{\text{lap}}\mathcal{L}_{\text{lap}} \\ &+ w_{\text{interp}}\mathcal{L}_{\text{interp}} + w_{\text{w}}\mathcal{L}_{\text{w}}\end{aligned} \tag{4.8}$$

- **3D vertex loss for clothing.** We define $\mathcal{L}_{3D}$ as L_1 loss between predicted and ground truth vertices

$$\mathcal{L}_{3D} = ||\mathbf{G}_{\text{P}} - \mathbf{G}_{\text{GT}}||_1. \tag{4.9}$$

- **3D vertex loss for shape under clothing.** For training f_w^B (ParserNet for the body), we use the input mesh skin as supervision for predicting personal details of the subject. We define a clothing class specific geodesic distance weighted loss term, as shown in equation (4.10), where $\mathbf{I}^s$ is the indicator matrix for skin region and $\mathbf{w}_{\text{geo}}$ is a vector containing the *sigmoid* of the geodesic distances from vertices to the boundary between skin and non-skin regions. The loss term is high when the prediction is far from the input mesh $\mathbf{M}$ for the visible skin region, and lower for the cloth region, with a smooth transition regulated by the geodesic term. Let $\text{abs}_{ij}(\cdot)$ denote an element-wise absolute value operator. Then the loss is computed as

$$\mathcal{L}_{\text{3D}}^{\text{body}} = \|\mathbf{w}_{\text{geo}}^T \cdot \text{abs}_{ij}(\mathbf{G}_{\text{P}}^s - \mathbf{I}^s\mathbf{M})\|_1. \tag{4.10}$$

- **Normal Loss.** We define $\mathcal{L}_{\text{norm}}$ as the difference in angle between ground truth face normal ($\mathbf{N}_{\text{GT}}^i$) and predicted face normal ($\mathbf{N}_P^i$).

$$\mathcal{L}_{\text{norm}} = \frac{1}{N_{\text{faces}}} \sum_i^{N_{\text{faces}}} (1 - (\mathbf{N}_{\text{GT},i})^T \mathbf{N}_{P,i}). \tag{4.11}$$

- **Laplacian smoothness term.** This enforces the Laplacian of the predicted clothing mesh to be close to the Laplacian of the ground truth mesh. Let $\mathbf{L}^g \in \mathbb{R}^{m_g \times m_g}$ be the graph Laplacian of the clothing mesh $\mathbf{G}_{\text{GT}}$, and $\Delta_{\text{init}} = \mathbf{L}^g\mathbf{G}_{\text{GT}} \in \mathbb{R}^{m_g \times 3}$ be the differential coordinates of the $\mathbf{G}_{\text{GT}}$, then we compute the Laplacian smoothness term for a predicted mesh $\mathbf{G}_{\text{P}}$ as:

$$\mathcal{L}_{\text{lap}} = ||\Delta_{\text{init}} - \mathbf{L}^g\mathbf{G}_{\text{P}}||_2. \tag{4.12}$$

- **Interpenetration loss.**Since minimizing per-vertex loss does not guarantee that the predicted clothing lies outside the body surface, we use the interpenetration loss term in equation (4.20) proposed in GarNet [GCS*19]. For every vertex $\mathbf{G}_{\text{P},j}$, we find the nearest vertex in the predicted body shape under clothing ($\mathbf{B}_i$) and define the body-clothing correspondences as $\mathcal{C}(\mathbf{B}, \mathbf{G}_{\text{P}})$. Let $\mathbf{N}_i$ be the normal of the i^{th} body vertex $\mathbf{B}_i$. If the predicted clothing vertex $\mathbf{G}_{\text{P},j}$ penetrates the body, and it is penalized with the following loss

$$\mathcal{L}_{\text{interp}} = \sum_{(i,j)\in\mathcal{C}(\mathbf{B},\mathbf{G}_{\text{P}})} \mathbb{1}_{d(\mathbf{G}_{\text{P},j},\mathbf{G}_{\text{GT},j})<d_{tol}} ReLU(-\mathbf{N}_i(\mathbf{G}_{\text{P},j} - \mathbf{B}_i))/m_g, \tag{4.13}$$

where notice that $\mathbb{1}_{d(\mathbf{G}_{\text{P},j},\mathbf{G}_{\text{GT},j})<d_{tol}}$ activates the loss when the distance between predicted clothing mesh vertices and ground truth mesh vertices is small *i.e.* $< d_{tol}$.

- **Weight regularizer.** To preserve the fine details when parsing the input mesh, we want the weights predicted by the network to be sparse and confined in a local neighborhood. Hence, we add a regularizer which penalizes large values for $\mathbf{W}_{ij}$ if the distance

between of $\mathbf{M}_j$ and the vertex $\mathbf{M}_k$ with largest weight $k = \arg\max_j \mathbf{W}_{ij}$ is large. Let $d(\cdot,\cdot)$ dennote Euclidean distance between vertices, then the regularizer equals:

$$\mathcal{L}_w = \sum_{i=1}^{m_g} \sum_{j \in \mathcal{N}_i} \mathbf{W}_{ij} d(\mathbf{M}_k, \mathbf{M}_j), \; k = \arg\max_j \mathbf{W}_{ij}. \tag{4.14}$$

Implementation Details: We implement f_w^θ and f_w^β networks with 2 fully connected and a linear output layer. We implement *ParserNet* f_w^U, f_w^L, f_w^B with 3 fully connected layers. We use neighborhood ($\mathcal{N}_i$) size of $|\mathcal{N}_i| = 50$, for our experiments. We first train the network for clothing classes that share the same clothing template and then fine-tune separately for each clothing class g. To speed up training for *ParserNet*, we train the network to predict $\mathbf{W}^g = \mathbf{I}^g$, where $\mathbf{I}^g$ is the indicator matrix for clothing class g, explained in Section 4.2.2. This initializes the network to parse the clothing by cutting out a part of the input mesh based on the constant per-clothing indicator matrix, shown in Figure 4.4.

We held out 40 scans for testing in each clothing class, which includes some cases with unseen subjects and some with unseen clothing size for seen subjects. For the pose-shape prediction network and *ParserNet*, we use a batch size of 8 and a learning rate of 0.0001.

4.5 Clothing Resizing Model

In this section, we present *SizerNet*, an encoder-decoder-based deep network for clothing resizing. An overview of the method is shown in Figure 4.2.

4.5.1 Sizing deformation model

We aim to edit the clothing mesh based on clothing size labels such as S, M, L, *etc*, to see the dressing effect of the clothing for a new size. For this task, we propose an encoder-decoder-based network, as shown in Figure 4.2(right). The network f_w^{enc}, encodes the clothing mesh $\mathbf{G}_{in}$ to a lower-dimensional latent code $\mathbf{x}_{\text{gar}} \in \mathbb{R}^d$, shown in Equation (4.15). We append $(\beta, \delta_{\text{in}}, \delta_{\text{out}})$ to the latent space, where $\delta_{\text{in}}, \delta_{\text{out}}$ are one-hot encodings of input and desired output sizing and β is the SMPL β parameter for underlying body shape.

$$\mathbf{x}_{\text{gar}} = f_w^{\text{enc}}(\mathbf{G}_{in}), \quad f_w^{\text{enc}}(.) : \mathbb{R}^{m_g \times 3} \rightarrow \mathbb{R}^d \tag{4.15}$$

The decoder network, $f_w^{\text{dec}}(.) : \mathbb{R}^{|\beta|} \times \mathbb{R}^d \times \mathbb{R}^{2|\delta|} \rightarrow \mathbb{R}^{m_g \times 3}$ predicts the displacement field $\mathbf{D}^g = f_w^{\text{dec}}(\beta, \mathbf{x}_{\text{gar}}, \delta_{\text{in}}, \delta_{\text{out}})$ on top on template. We obtain the output clothing $\mathbf{G}_{out}$ in the new desired size δ_{out} using Equation (4.4).

4.5.2 Training and Implementation Details

Loss Functions: We train *SizerNet* with training loss given by Equation (4.16) respectively, where w_{3D}, w_{norm}, w_{lap} and w_{interp} are weights for the loss on vertices, normal, Laplacian, and interpenetration term respectively. We explain each of the loss terms in this section.

$$\mathcal{L}_{\text{sizer}} = w_{\text{3D}}\mathcal{L}_{\text{3D}} + w_{\text{norm}}\mathcal{L}_{\text{norm}} + w_{\text{lap}}\mathcal{L}_{\text{lap}} + w_{\text{interp}}\mathcal{L}_{\text{interp}} \tag{4.16}$$

We now define the individual loss used in training the network. For more details, refer to Section 4.5.2.

- **3D vertex loss for clothing.** We define $\mathcal{L}_{\text{3D}}$ as L_1 loss between predicted and ground truth vertices
$$\mathcal{L}_{\text{3D}} = ||\mathbf{G}_{\text{P}} - \mathbf{G}_{\text{GT}}||_1. \tag{4.17}$$
- **Normal Loss.** We define $\mathcal{L}_{\text{norm}}$ as the difference in angle between ground truth face normal ($\mathbf{N}^i_{\text{GT}}$) and predicted face normal ($\mathbf{N}^i_P$).
$$\mathcal{L}_{\text{norm}} = \frac{1}{N_{\text{faces}}} \sum_{i}^{N_{\text{faces}}} (1 - (\mathbf{N}_{\text{GT},i})^T \mathbf{N}_{P,i}). \tag{4.18}$$
- **Laplacian smoothness term.** This enforces the Laplacian of predicted clothing mesh to be close to the Laplacian of ground truth mesh and is given by:
$$\mathcal{L}_{\text{lap}} = ||\Delta_{\text{init}} - \mathbf{L}^g \mathbf{G}_{\text{P}}||_2. \tag{4.19}$$
- **Interpenetration loss.** This is exactly similar to the loss implemented for ParserNet. For every vertex $\mathbf{G}_{\text{P},j}$, we find the nearest vertex in the fixed body shape under clothing ($\mathbf{B}_i$) and define the body-clothing correspondences as $\mathcal{C}(\mathbf{B}, \mathbf{G}_{\text{P}})$. Let $\mathbf{N}_i$ be the normal of the i^{th} body vertex $\mathbf{B}_i$. If the predicted clothing vertex $\mathbf{G}_{\text{P},j}$ penetrates the body, it is penalized with the following loss
$$\mathcal{L}_{\text{interp}} = \sum_{(i,j)\in\mathcal{C}(\mathbf{B},\mathbf{G}_{\text{P}})} \mathbb{1}_{d(\mathbf{G}_{\text{P},j},\mathbf{G}_{\text{GT},j})<d_{tol}} ReLU(-\mathbf{N}_i(\mathbf{G}_{\text{P},j} - \mathbf{B}_i))/m_g, \tag{4.20}$$
where notice that $\mathbb{1}_{d(\mathbf{G}_{\text{P},j},\mathbf{G}_{\text{GT},j})<d_{tol}}$ activates the loss when the distance between predicted clothing mesh vertices and ground truth mesh vertices is small *i.e.* $< d_{tol}$.

Implementation Details: For *SizerNet*, we use $d = 30$ and we implement f_w^{enc}, f_w^{dec} with fully connected layers and skip connections between encoder and decoder network. We held out 40 scans for testing in each clothing class, which includes some cases with unseen subjects and some with unseen clothing size for seen subjects. We use a batch size of 8 and a learning rate of 0.0001.

4.6 Experiments and Results

In this section, we first show the results of *3D Clothing Parsing* in Section 4.6.1, followed by results of *Clothing Resizing* in Section 4.6.2. Since *SizerNet* learns both body shape and clothing size dependent deformation, it can also be used for draping clothing to new body types, and we show the results in Section 4.6.3.

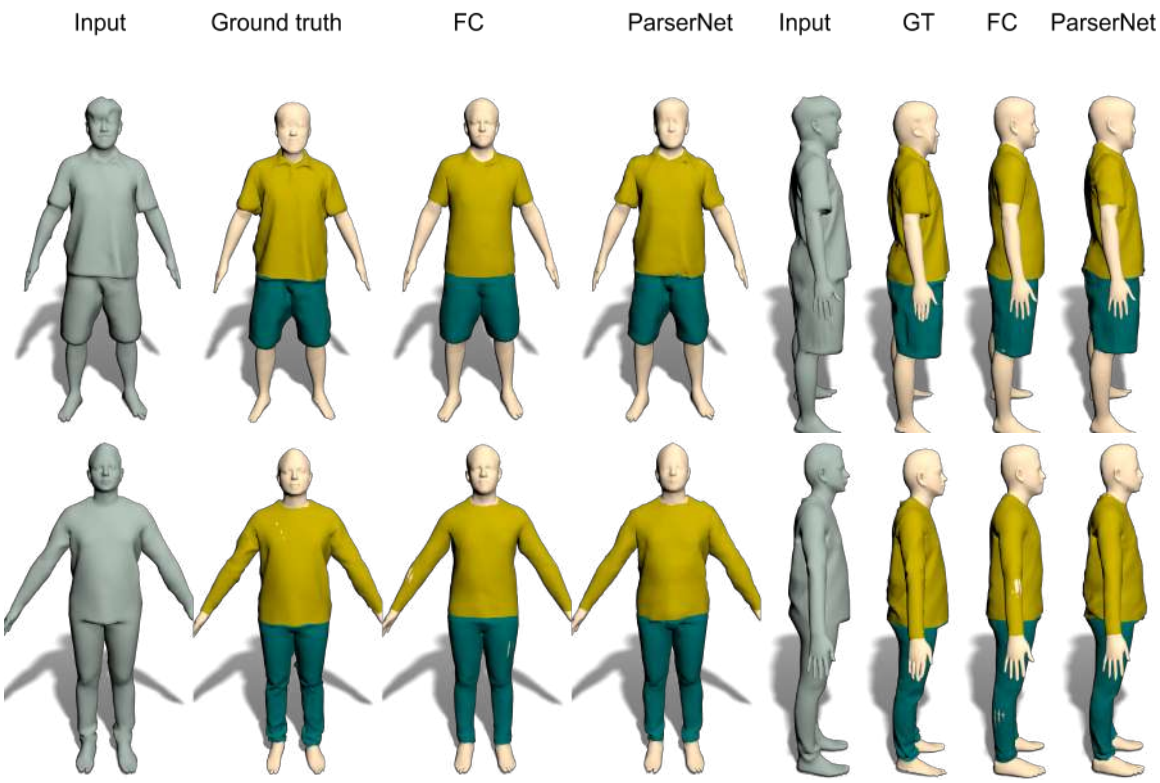


Fig. 4.5 **Comparison of *ParserNet* with FC-Naive:** We observe that FC-Naive is not able to preserve the fine details and wrinkles present in the single-layered input mesh. Our ParserNet formulation, which is based on a sparse weighted sum of input vertices, preserves the details and is also able to better explain loose clothing than FC-Naive.

4.6.1 3D Clothing Parsing and Shape under Clothing

As discussed in Section 4.4.1, we propose to parse clothing from a single-layered mesh by deforming the vertices of a template mesh by using a convex combination of the vertices of the input single-layered mesh. If we parse clothing using this approach, instead of directly predicting the location of clothing vertices, we preserve more details in the parsed layer. To validate the choice of parsing the clothing using a sparse regressor matrix ($\mathbf{W}$), we compare the results of *ParserNet* with two baseline approaches:

- **Linear Model**: A linearized version of *ParserNet* implemented with LASSO. Instead of predicting the weight matrix using a neural network, we solve a linear problem using LASSO.
- **FC-Naive**: A naive FC network, which has the same architecture as *ParserNet*. However, instead of predicting the weight matrix ($\mathbf{W}$), the FC network directly predicts the deformation ($\mathbf{D}^g$) from the clothing template ($T^g(\beta,\theta,\mathbf{0})$) for a given input ($\mathbf{M}$).

We compare the per-vertex error of *ParserNet* with the aforementioned baselines in Table 4.4. We compare the qualitative results of *ParserNet* with the naive FC network in

Clothing	Linear Model	FC-Naive	ParserNet	Clothing	Linear Model	FC-Naive	ParserNet
Polo	32.21	17.25	**14.33**	Shorts	29.78	20.12	**16.07**
Shirt	27.63	19.35	**14.56**	Pants	34.82	18.20	**17.24**
Vest	28.17	18.56	**15.89**	Coat	41.27	22.19	**15.34**
Hoodies	37.34	23.69	**15.76**	Shorts2	31.38	23.45	**16.23**
T-Shirt	26.94	15.98	**13.77**				

Table 4.4 **Average per-vertex error V_{err} for *3D Clothing Parsing*:** We compare the parsed clothing with ground-truth clothing meshes registered to a predefined common template across each clothing class. We observe that ParserNet achieves better results than Linear Model and FC-Naive. All the numbers are reported in *mm*.

figure 4.5, which shows that *ParserNet* can produce details, fine wrinkles, and large clothing deformations. Our approach leverages a sparse linear formulation, where each clothing vertex is expressed as a weighted combination of input mesh vertices via a learned sparse matrix. This representation introduces two key advantages. First, the sparsity constraint enforces local support and preservation of local details, such that each clothing vertex depends on a limited set of source vertices. Moreover, since the final output is a weighted copy of the input observation, it preserves the small and large details present in the input mesh, and hence it is able to explain very loose clothing, large folds, and fine wrinkles. This is not possible with a naive FC network. Since *ParserNet* reconstructs the output clothing mesh as a localized sparse weighted sum of input vertex locations, it preserves the geometry details present in the input mesh. However, in the case of a naive FC network, the predicted displacement field ($\mathbf{D}^g$) is smooth and does not explain large deformations. Hence, a naive FC network is unable to predict loose clothing and fails to preserve fine details.

The *ParserNet* formulation is flexible and works for diverse kinds of clothing classes, as long as they follow body topology. We show results of *ParserNet* for more clothing classes in Figure 4.6. We observe that our proposed formulation works for diverse types of clothing, such as t-shirts, hoodies, and vests, while preserving details in the input single-layered mesh. For more results, please refer to the full method results shown in Section 4.6.2.1.

4.6.2 3D Clothing Resizing

Editing clothing meshes based on clothing size label is an unexplored and challenging problem. As a result, there are no well-defined quantitative metrics and baselines. We introduce two quantitative metrics, namely change in mesh *surface area* (A_{err}) and *per-vertex error* (V_{err}) for evaluating the resizing task. *Surface area* accounts for the scale of a clothing, which only changes with the clothing size, and *per-vertex error* accounts for details and folds created due to the underlying body shape and looseness/tightness of the clothing. Moreover, subtle changes in clothing shape with respect to size are difficult to evaluate quantitatively and are not uniform throughout the surface of clothing. Hence, we use heat map visualizations for qualitative analysis of the results.

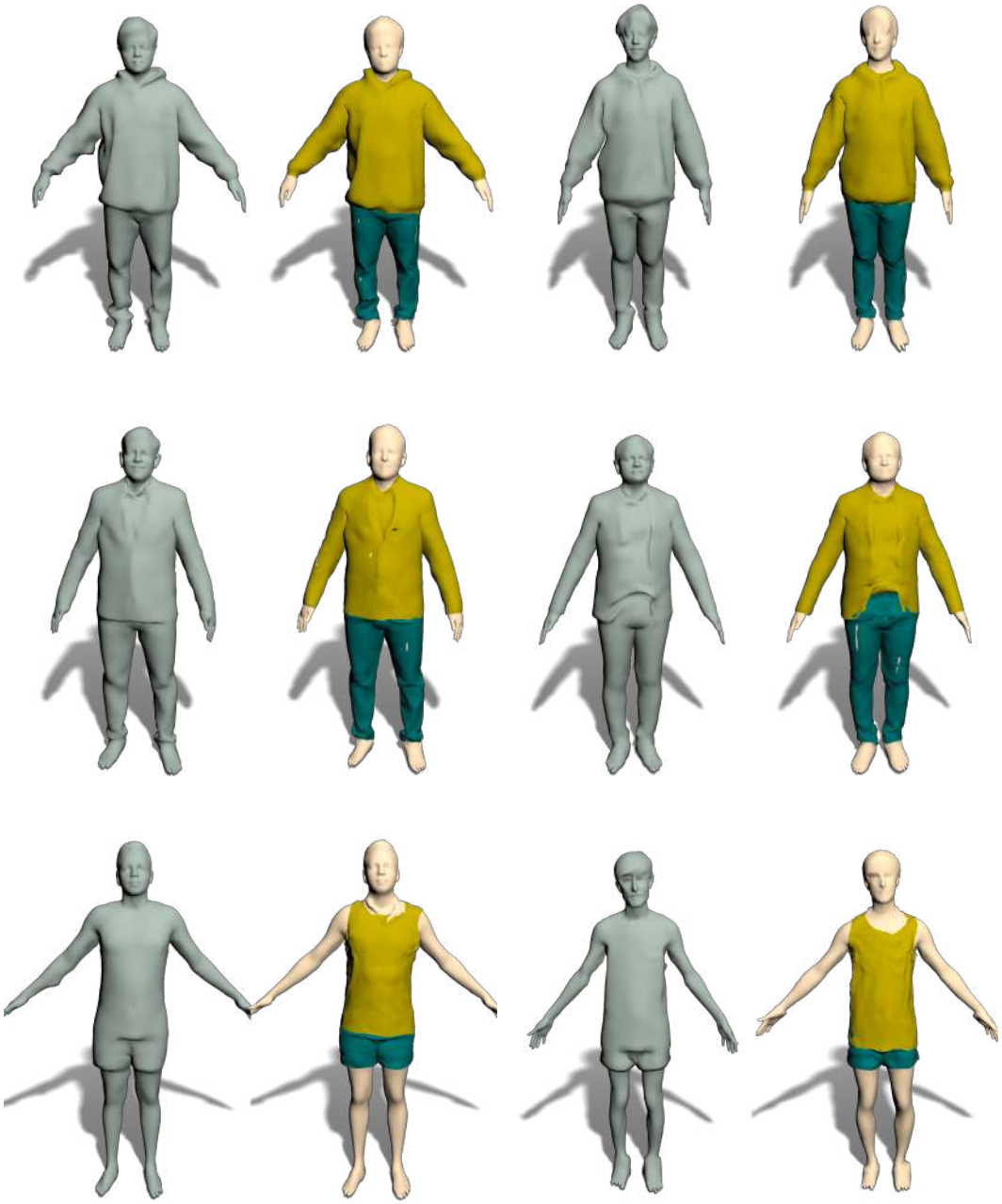


Fig. 4.6 ***3D Clothing Parsing* for diverse clothing from input single-layered mesh**. Using ParserNet, one can parse a multi-layered representation of clothing and the body underneath. We present results of ParserNet across different clothing classes and demonstrate that it generalizes well to various clothing styles while preserving fine details.

Clothing	Error-Margin		Average-Pred		Linear Scaling		Ours	
	V_{err}	A_{err}	V_{err}	A_{err}	V_{err}	A_{err}	V_{err}	A_{err}
Polo T-shirt	33.25	24.56	23.86	3.63	35.05	8.45	**16.42**	**1.79**
Shirt	36.52	19.57	21.95	2.76	34.53	7.01	**15.54**	**1.41**
Shorts	43.21	27.21	24.79	5.41	35.77	4.99	**16.71**	**2.38**
Pants	30.83	15.15	21.54	4.73	38.16	7.13	**19.26**	**2.43**
Vest	35.46	15.41	25.81	2.98	29.63	7.37	**16.03**	**1.84**
Coat	39.43	18.29	23.13	3.16	43.64	9.93	**15.37**	**1.75**
Hoodies	40.09	16.65	26.39	4.74	41.42	8.78	**15.89**	**1.46**
T-shirt	32.23	21.66	–	–	34.59	7.83	**14.98**	**1.39**
Shorts2	27.35	24.98	19.82	4.71	36.32	5.69	**16.83**	**2.21**

Table 4.5 **Average per-vertex error (V_{err} in mm) and surface area error (A_{err} in %) for clothing resizing:** We compare the aforementioned metrics between clothing predicted using SizerNet and the ground-truth clothing meshes, which are registered to a common template for each clothing class. We evaluate our model against the ground-truth error margin across all sizes, the average prediction, and a linear scaling model. These results suggest that clothing resizing is a highly complex task and requires data-driven deep models to learn the underlying deformations.

Since there is no other existing work for the clothing resizing task to compare with, we evaluate our method against the following three baselines.

1. **Error margin** in data: We define error margin as the change in *per-vertex location* (V_{err}) and *surface area* (A_{err}) between clothing of two consecutive sizes for a subject in the dataset. Our model should ideally produce an error smaller than this margin.

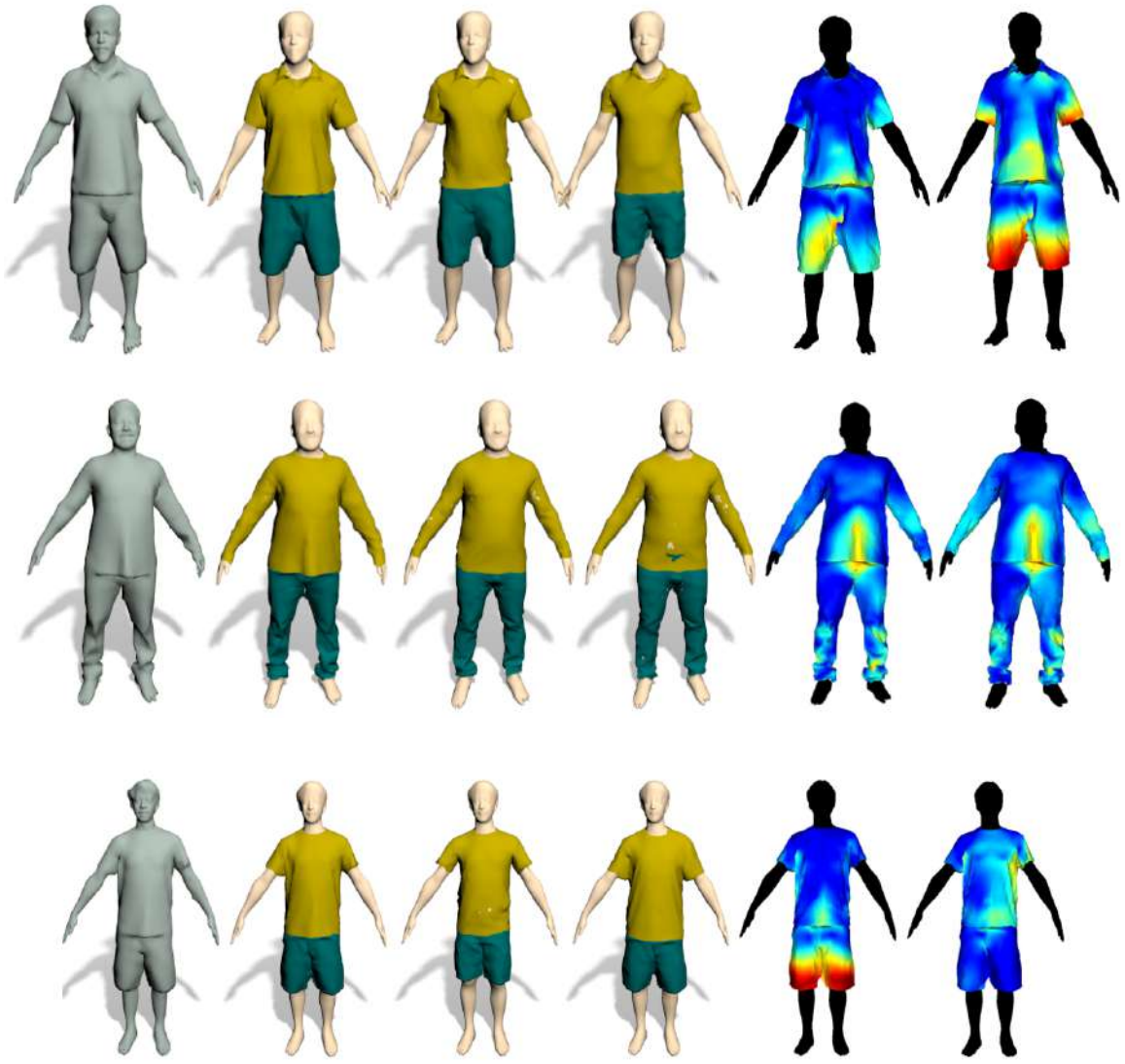
2. **Average prediction**: For every subject in the dataset, we create the average clothing (G_{avg}) by averaging over all the available sizes for a subject.

3. **Linear scaling + Alignment**: We linearly scale the clothing mesh, according to the desired size label, and then align the clothing to the underlying body.

Table 4.5 shows the errors for each experiment. *SizerNet* yields lower errors compared to the linear scaling method, reflecting the need to model the non-linear relationship between clothing shape, underlying body shape, and clothing size. We also observe that network predictions yield lower error compared to average clothing prediction, which suggests that the model is learning to account for size variation, even though the differences in the ground truth itself are subtle. We present the results of *SizerNet* for common clothing classes in Table 4.5.

4.6.2.1 Resizing Clothing from a Single Mesh of Human in Clothing

As shown in Figure 4.2, our whole pipeline leverages a tandem of ParserNet and SizerNet to enable practical applications in 3D clothing editing. Given a single-layered clothing



(a) Medium(input, parsed), Small, Large.



Fig. 4.7 **ParsetNet+SizerNet module for extracting and resizing clothing**(*Left to Right*) Input single mesh, parsed multi-layer mesh from ParserNet, Resized clothing in two subsequent smaller sizes, heatmap of change in per-vertex error on original parsed clothing for two new sizes.

mesh registered to a body model such as SMPL, we can visualize and manipulate the sizing effects on individual clothing items. First, we apply our pose-shape prediction networks alongside the Clothing+Body ParserNets to generate a disentangled, multi-layered mesh representation. This serves as input to SizerNet, which enables precise resizing of each clothing item independently.

In this section, we show results across diverse clothing styles and provide detailed heatmap analyses to highlight the distinct contributions of both ParserNet and SizerNet. The entire workflow can be built upon publicly available SMPL registration pipelines, seamlessly integrating with our networks to form a robust editing framework.

Such fine-grained clothing manipulation would not be possible without the synergy of these two models. In Figure 4.7, we showcase parsing of everyday clothing items followed by resizing to progressively smaller sizes. We further visualize the per-vertex difference heatmaps, which reveal that resizing involves a blend of both global transformations and localized deformations, driven by the interaction between clothing resizing and body draping effects.

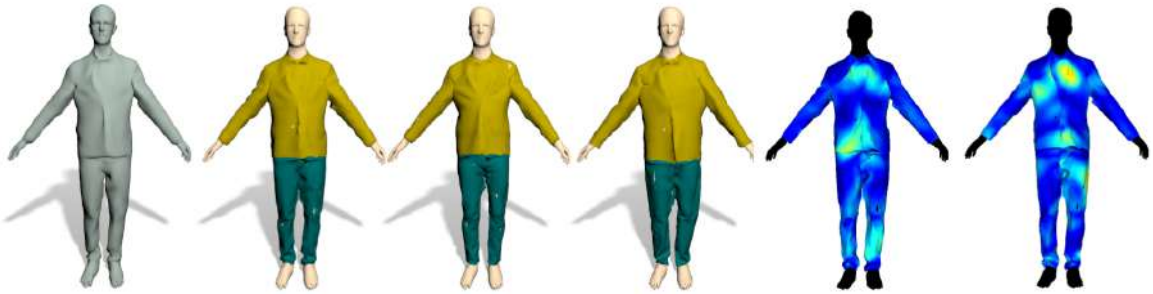
We also present the results of *SizerNet* for various clothing categories, including hoodies, coats, vests, and others. Figure 4.8 illustrates the resized clothing for a given input mesh, along with a heatmap showing the per-vertex displacement relative to the input (parsed) clothing. From Figure 4.8, we observe that the changes are most prominent near the clothing boundaries, where dimensions such as sleeve length and hem length vary with size.

For clothing items like shorts, T-shirts, and hoodies, we observe more noticeable changes compared to clothing like shirts and coats. This is due to the inherent style of these clothing items–for instance, coats and shirts often follow a slim-fit design, which results in less variation with changes in size.

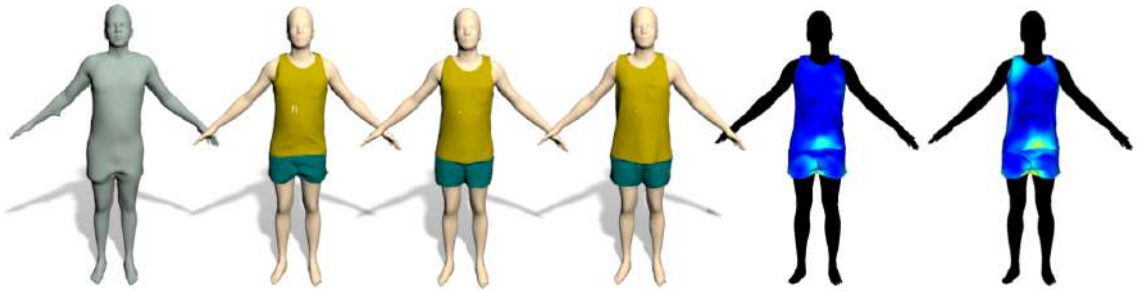
We present additional qualitative results demonstrating the effectiveness of our clothing extraction and resizing pipeline. Specifically, we use ParserNet to parse or extract individual clothing items from a single human mesh input. These clothes are then resized to simulate variations in fit, including both one size smaller and one or two sizes larger, conditioned on the discrete size labels. We show results for all clothing classes available in the SIZER dataset in Figure 4.9, Figure 4.11 Figure 4.12, Figure 4.13, and Figure 4.14. To demonstrate that our model can resize clothing in either direction, we provide examples of downsizing results as well. Some results contain upsizing up to 2 size labels. We observe that SizerNet accurately predicts the resulting deformation of the clothing geometry under various size modifications. It effectively captures fine-grained, spatially varying, and non-uniform deformation patterns that occur naturally during garment scaling. This highlights the model's ability to preserve realistic garment behavior.



(a) Large(input, parsed), Medium, Small.



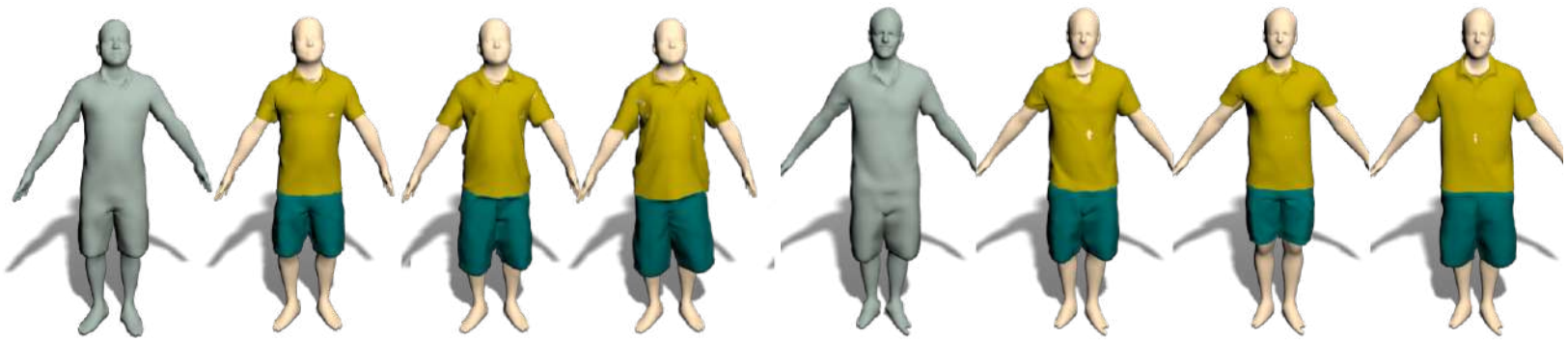
(b) Large(input, parsed), Medium, X-Large.



(c) Small(input, parsed), Medium, Large.

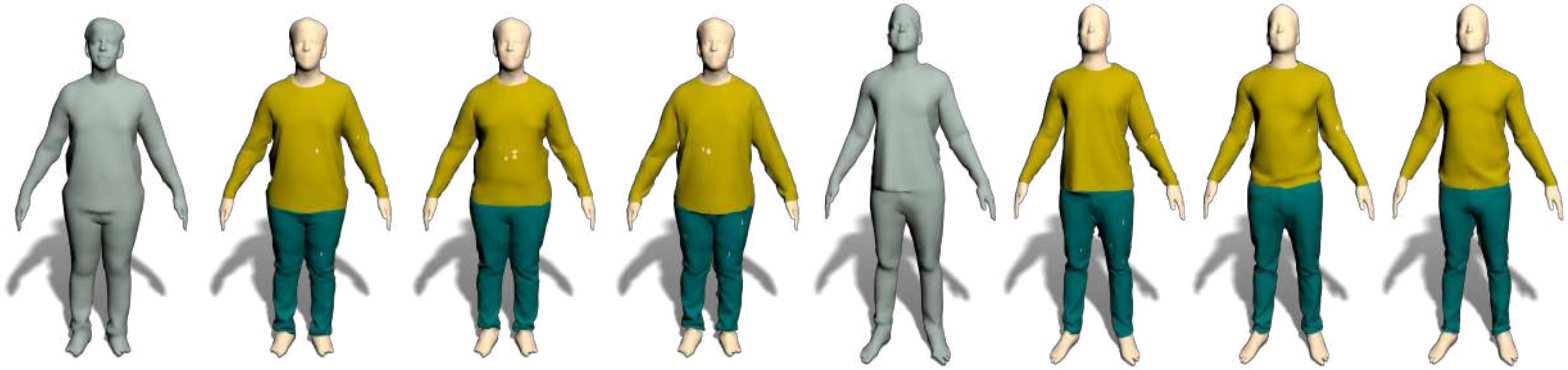


Fig. 4.8 **ParsetNet+SizerNet module for extracting and resizing clothing**(*Left to Right*): Input single mesh, parsed multi-layer mesh from ParserNet, Resized clothing in two new sizes, heatmap of change in per-vertex error on original parsed clothing for two new sizes.



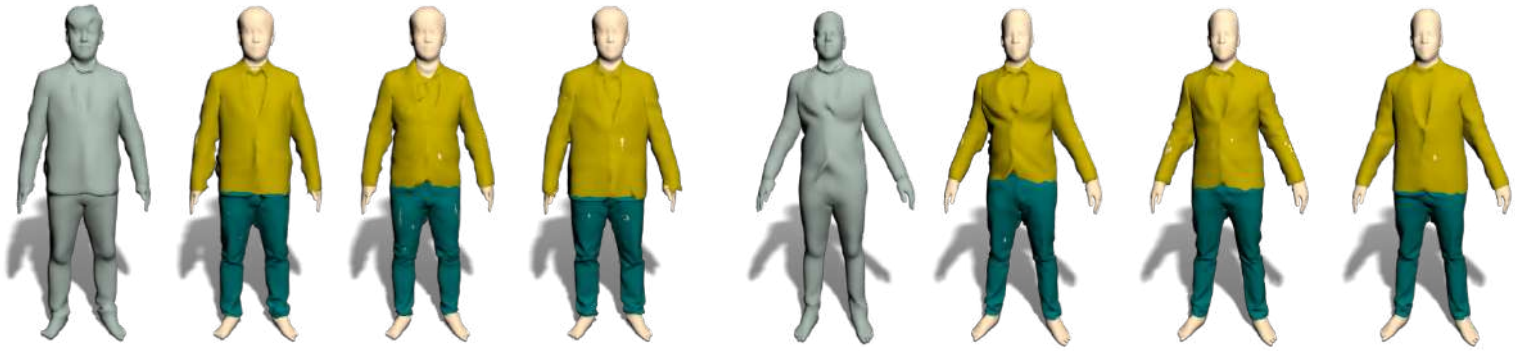
(a) Small(input, parsed), Medium, Large (b) Medium(input, parsed), Small, Large

Fig. 4.9 **Results of *ParserNet* + *SizerNet* for T-Shirt and Shorts**, where we parse the clothing from a input single mesh and change the size of the clothing to visualise dressing effect.



(a) Large(input, parsed), Medium, XLarge (b) XLarge(input,parse), Large, Medium

Fig. 4.10 **Results of *ParserNet* + *SizerNet* for Shirt and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect.



(a) Large(input, parsed), Medium, XL.

(b) Medium(input, parsed), Large, XL.

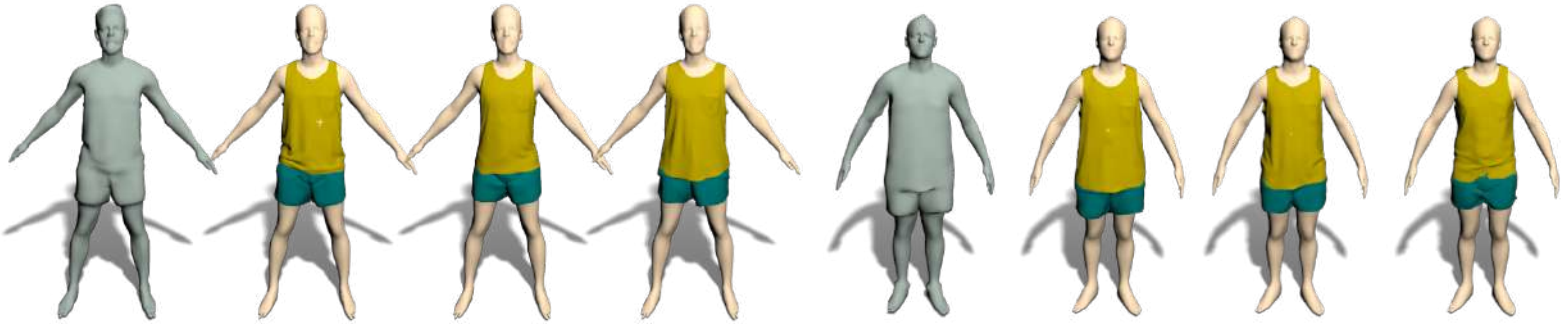
Fig. 4.11 **Results of *ParserNet* + *SizerNet* for Coat and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect.



(a) Medium(input, parsed), Small, Large.

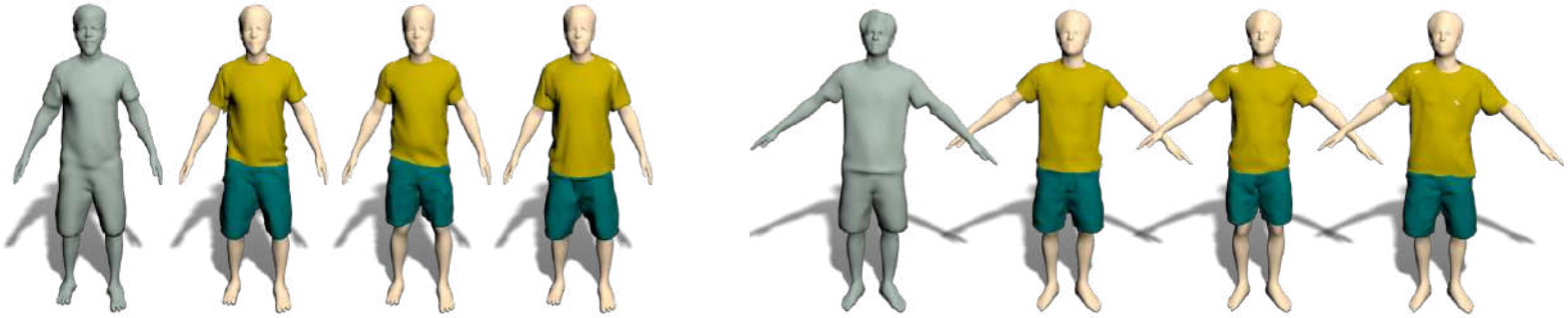
(b) Medium(input, parsed), Small, Large.

Fig. 4.12 **Results of *ParserNet* + *SizerNet* for Hoodies and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect.



(a) Medium(input, parsed), Small, Large. (b) Large(input, parsed), Medium, Small.

Fig. 4.13 **Results of *ParserNet* + *SizerNet* for Vest and Shorts**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect.



(a) Medium(input, parsed), Small, Large. (b) Small(input, parsed), Medium, Large.

Fig. 4.14 **Results of *ParserNet* + *SizerNet* for TShirt and Shorts**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect.

4.6.3 Draping on Unseen Body

We can also drape a clothing mesh on a body and change its shape. For this, we take clothing mesh as input to SizerNet which is encoded (f_w^{enc}) into lower-dimensional latent vector and we append β of a new body and input and desired size labels ($\delta_{\text{in}}, \delta_{\text{out}}$) in the latent space. The decoder (f_w^{dec})) then predicts displacement from the clothing template and drapes the input clothing mesh on this new body. Figure 4.16 shows the results of draping clothing on various people, without changing the absolute size of the clothing. We present the results for shirt, pants, t-shirt, and shorts.

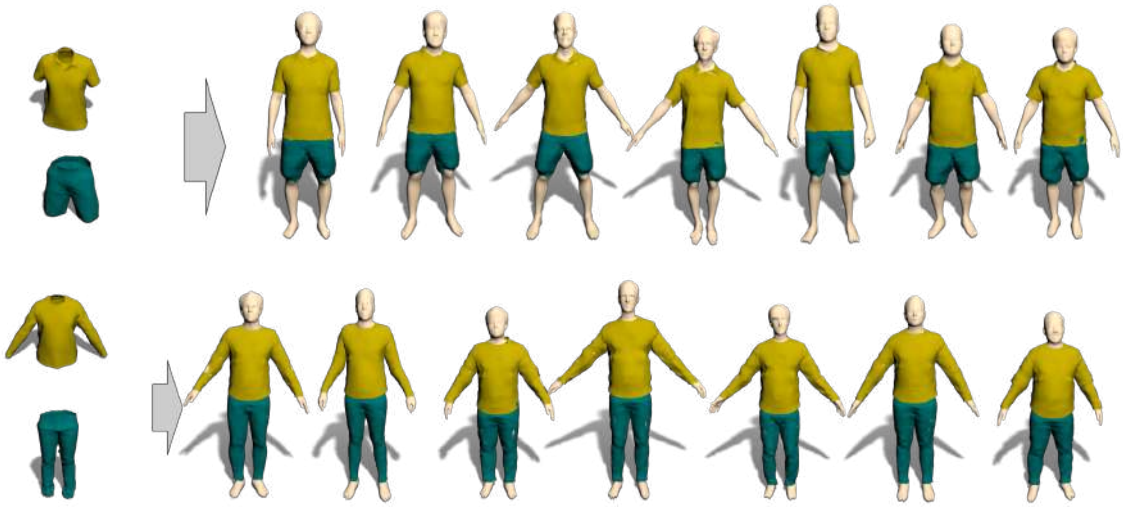
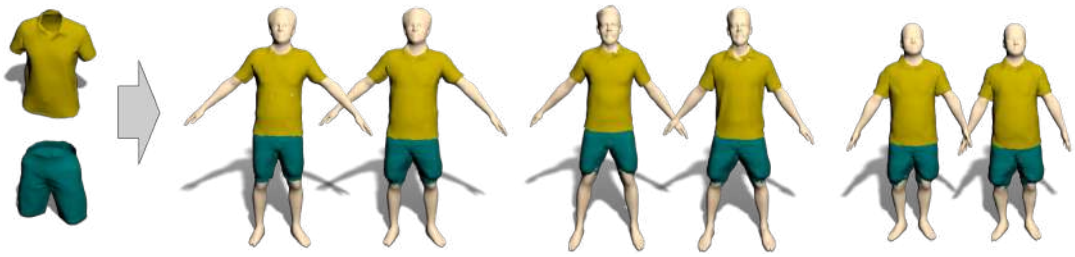


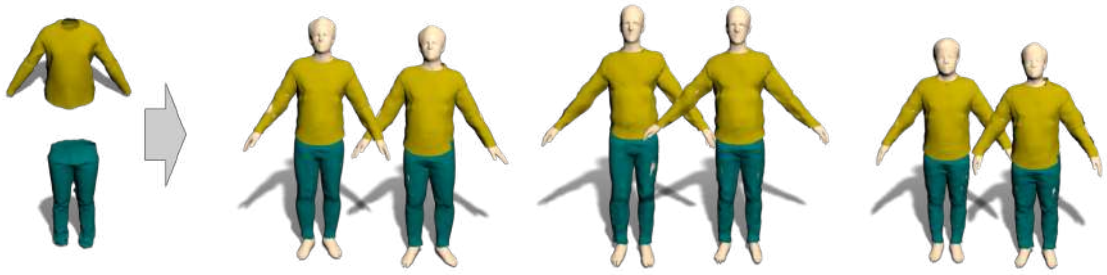
Fig. 4.15 **Draping clothes on different body shape:** Given an input clothing in a fixed size, SizerNet can also drape the clothing to different body shapes to see the effect of fixed clothing size on different body shapes.

4.6.4 Ablation study

We experiment with the significance of each loss function introduced in the method section. We present our result with and without using the interpenetration term in the training, and as shown in Figure 4.17(a,b), the network produces fewer intersections when trained with $\mathcal{L}_{\text{interp}}$. Quantitatively, in ParserNet, 6.9% vertices of the predicted clothing mesh intersects with the underlying body, which reduces to 3.5%, when trained with $\mathcal{L}_{\text{interp}}$. This difference is more prominent in polo t-shirts, t-shirts, and hoodies, and less in the case of shirt class. We also present our results with and without the Laplacian loss term, and we observe a smooth surface in the prediction when the Laplacian loss is present, especially near the shoulder region. However, in the absence of this surface regulariser, the model produces some noisy details, as shown in Figure 4.18.



(a) Input (Medium) (b) Small, Medium (c) Small, Large (d) Small, Medium.



(b) Input (Large) (b) Medium, Large (c) Large, Medium (d) Small, Medium.

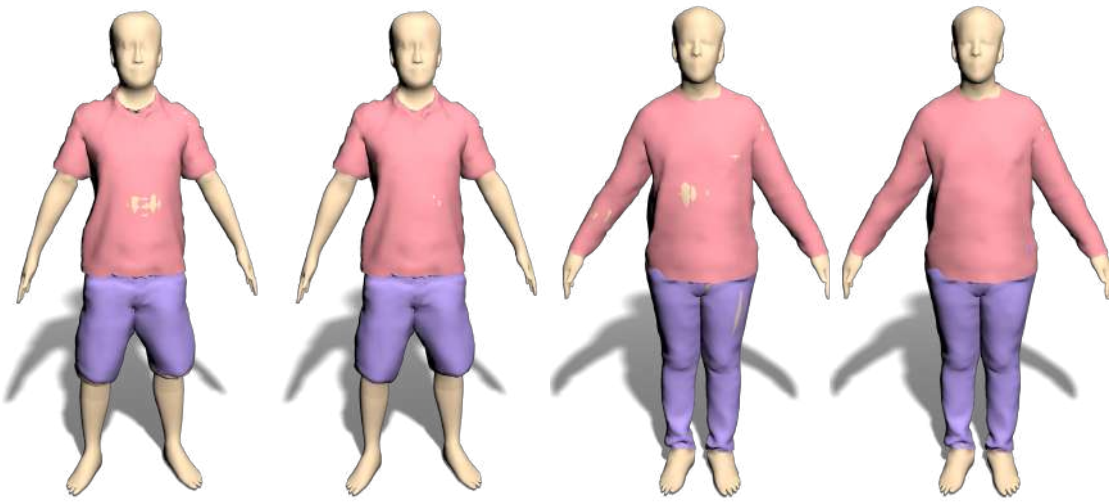
Fig. 4.16 **Draping clothes on different body shapes and modifying the clothing size:** Given an input clothing in a fixed size, SizerNet can also drape the clothing to different body shapes and further modify the clothing size.

4.7 Conclusion

We introduce *SIZER*, a clothing size variation dataset and model, which is the first real dataset to capture clothing size variation on different subjects. We introduce *ParserNet*: a 3D clothing parsing network and *SizerNet*: a size-sensitive clothing model. With this method, one can change the single mesh registration to multi-layer meshes of clothing and body shape under clothing, without the need for scan segmentation, and can use the result for animation, virtual try-on, etc. (*SizerNet*) can drape a person with clothing in different sizes. We released the model, dataset, and code to stimulate research in the direction of 3D clothing parsing, segmentation, resizing, and predicting body shape under clothing. Further, the novel sparse

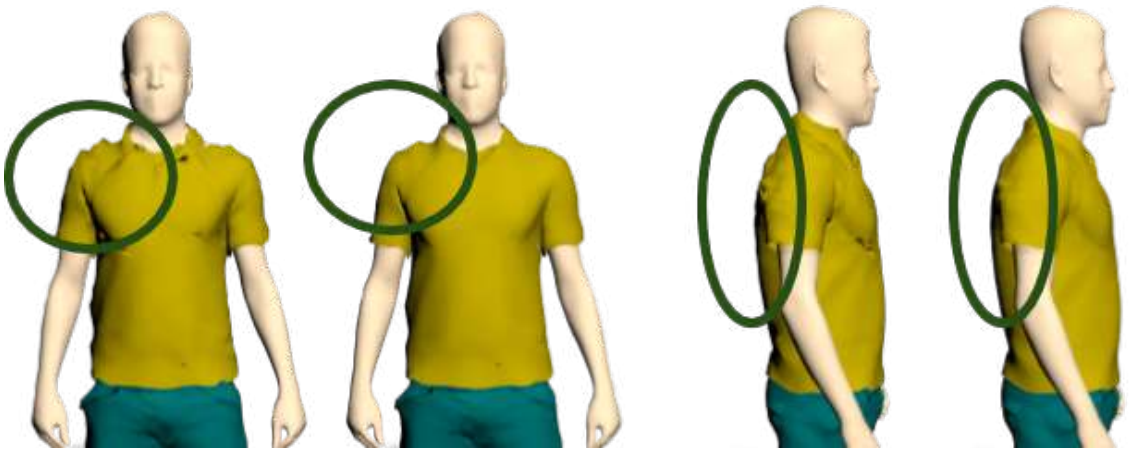
Loss	**Shirt**	**Polo-T-Shirt**	**ShortPants**	**Pants**
$\mathcal{L}_{\text{interp}} + \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{3D}}$	14.56	14.33	16.07	17.24
$\mathcal{L}_{\text{interp}} + \mathcal{L}_{\text{3D}}$	14.41	14.34	16.08	17.09
$\mathcal{L}_{\text{3D}}$	14.23	14.27	15.98	16.93

Table 4.6 **ParserNet Ablation:** Impact of surface regulariser and intersection loss in ParserNet predictions.



(a) (b)

Fig. 4.17 **Interpenetration loss in ParserNet:** Results of ParserNet trained without and with $\mathcal{L}_{\text{interp}}$. We observe that with interpenetrtation loss, the parsed clothing has less intersection with the underlying body.



(a)

Fig. 4.18 **Laplacian loss in ParserNet:** Results of ParserNet trained without and with $\mathcal{L}_{\text{lap}}$. We observe that with laplacian loss, the parsed clothing has more realsitic and smooth deformation.

linear formulation of ParserNet enables localized, interpretable mappings from input mesh to clothing by directly combining source vertices rather than predicting new ones.

4.7.1 Limitations and Discussions

While our proposed framework combining ParserNet and SizerNet represents a significant step toward size-aware multi-layered clothing modeling, several limitations remain. We discuss different aspects of challenges in the following paragraphs.

Data Processing and Registration: Our dataset comprises raw 3D scans that were aligned to SMPL using SMPL+D and SMPL+G representations. However, the registration process, especially for loose clothing such as coats or long skirts, is a challenging task in itself and also loses significant high-frequency details due to the limited and fixed resolution of the predefined template. The limitation stems from the inherent mismatch between template-based models (e.g., SMPL+G) and the diversity of real-world clothing, particularly when clothing deviates from the supported submesh topology or includes complex accessories (e.g., hoods, skirts).

Diversity and Pose Limitations: Despite the SIZER dataset covering 100 subjects and approximately 10 clothing classes, it is limited to static scans in near-canonical A-poses, which constrains the capture of pose-dependent draping effects. As clothing fit and wrinkle formation are highly pose-sensitive, our resizing and parsing models cannot yet model dynamic effects (as explored in DeepWrinkles [LCT18b]). This restricts downstream applications, such as motion synthesis or realistic try-on under animation.

Mesh-based Representation: The reliance on mesh-based representations offers strong compatibility with animation pipelines and supports per-vertex interpretability, which is beneficial for downstream tasks such as rendering, simulation, and editing. However, this approach imposes certain limitations on generalization: (1) the fixed topology of meshes makes it challenging to represent clothing with complex structures, such as skirts with slits or asymmetrical designs; and (2) the representation is inherently non-differentiable with respect to topological changes, limiting its applicability in learning-based frameworks compared to implicit or point-based representations. Reliance on predefined clothing templates restricts the model's extensibility to unseen or diverse clothing styles.

Motivated by these limitations, we explore alternative representations for clothed humans in subsequent chapters. Specifically, we explore neural implicit-based representation in Chapter 5 and point-cloud based representation in Chapter 6, where we analyze the trade-offs in scalability, expressiveness, and applicability offered by mesh-based, implicit, and point-based approaches.

4.7.2 Ethical Considerations

Clothed human modeling, particularly from real-world scans, raises important ethical considerations, especially around privacy and consent. The SIZER dataset comprises high-resolution 3D scans of real subjects. Although the subjects are clothed, the reconstructions often enable approximate inference of body shape under clothing, which can be considered

sensitive biometric data. We carefully designed a data collection protocol that allowed participants to choose their level of data sharing. All these preferences and associated guidelines are documented and included in the data release. Some key protocols and guidelines to avoid misuse of data are as follows:

- All data in the SIZER dataset were collected with explicit, informed consent from participants, including a clear explanation of how their data would be used. Participants were given the option to consent to full data release or to restrict the use of their scans for publication visuals.
- Distribution is restricted to research purposes only, and applicants are required to agree to usage terms that prohibit misuse (e.g., commercialization or identification).
- ParserNet and SizerNet only operate on registered meshes and do not require any RGB input, avoiding additional privacy concerns stemming from facial or skin texture analysis.

4.7.3 Future Work and Impact

This work lays the foundation for data-driven 3D modeling of size-dependent clothing deformation, and its impact is already visible in the follow-up works that extend or complement it. At the time of publication, most works addressed either pose-dependent [LCT18b, GRH*12] or shape-conditioned clothing deformation [GRH*12, PLPM20]. SizerNet was the first 3D model to incorporate size-sensitive clothing. Beyond addressing the limitations outlined in Section 4.7.1, we also consider broader future directions. Some of which have already been explored by the community, while others remain open challenges.

Why 3D Resizing Remains Challenging? Despite recent advances, 3D clothing resizing remains an open problem. Due to the high cost of capturing 3D clothing size variation data and the versatility required for clothing resizing in the real world, this area has not yet received sufficient attention. Image-based resizing [CCSC23] often fails to capture accurate draping effects due to the complex interaction between the body and the clothing. We argue that with the emergence of large-scale, sewing pattern–based clothing datasets [KKK*24, KSH23] and foundational models, 3D clothing resizing can flourish and become a powerful and cost-effective tool for size-based virtual try-on.

Layered Clothing Representation. With ParserNet, we proposed the first learning based method to obtain a multi-layered mesh from a single SMPL registration, a paradigm adopted by concurrent and follow-up works like BCNet [JZH*20], TightCap [CPW*21], GALA [KKSJ24], DrapeNet [DLLG*23], which all explore layered clothing representation. These models often use SIZER as a benchmark dataset for evaluating the extraction of layered clothing.

Impact of SIZER Dataset. The SIZER dataset remains a cornerstone benchmark for tasks including size-aware virtual try-on, clothing segmentation and parsing [ATO*24, MMC22], shape under clothing [DLLG*23], 3d reconstruction [XXMPM24], clothing fitting [CPA*21a, CJS*22, LGF23], and editing [LGF23, HYZ24].

In summary, this chapter introduces a novel framework that not only addresses a fundamental problem but also enables a new class of applications in e-commerce, avatar generation, and virtual try-on. We anticipate that our methods and dataset will serve as pivotal tools for future research in this domain.

5

NeuralGIF for Animating People in Clothing

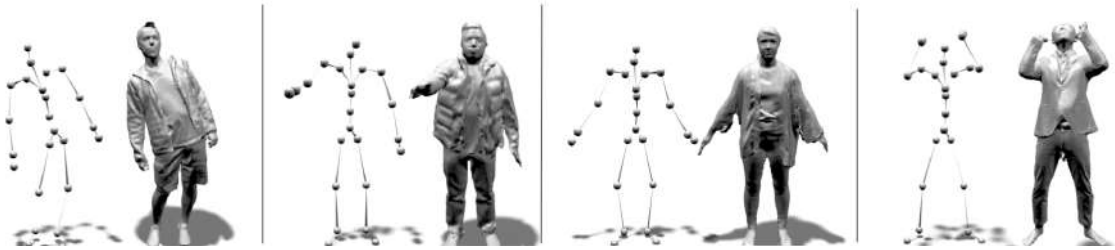


Fig. 5.1 We present **Neural Generalized Implicit Functions** (Neural-GIF), to animate people in clothing as a function of body pose. Neural-GIF learns directly from scans, models complex clothing and produces pose-dependent details for realistic animation. We show for four different characters with the query input pose on the left (illustrated with a skeleton) and our output animation on the right.

Animating *people in clothing* has wide applications in virtual avatar-driven works. Past works have modeled people in clothing using mesh-based representation [MYR*20, SOC19, PLPM20]. Driving the animation of complicated clothing using mesh-based methods has been suboptimal due to the limited resolution of meshes, the fixed predefined topology of meshes, or the complexity resulting from extensive data processing/generation requirements. With advances in more flexible neural implicit base representations, high-quality and detailed human reconstruction has started to emerge [SHN*19, SSSJ20]. In this direction, we present Neural Generalized Implicit Functions (Neural-GIF) to animate people in clothing as a function of body pose using the neural implicit representation. Given a sequence of scans of a subject in various poses, we learn to animate the character for new poses.

As discussed, existing methods have relied on template-based representations of the human body (or clothing). However, such models usually have fixed and limited resolution, require difficult data pre-processing steps, and cannot be used with complex clothing, *e.g.*, for clothing shown in Figure 5.1. However, template-based methods are easy to use and provide an efficient solution for animation. We draw inspiration from template-based methods, which factorize motion into articulation and non-rigid deformation, but generalize this concept for implicit shape learning to obtain a more flexible model. We learn to map every point in space to a canonical space, where a learned deformation field is applied to model non-rigid effects before evaluating the signed distance field.

Our formulation enables the learning of complex and non-rigid deformations of clothing and soft tissue without requiring template registration, as is common in current approaches.

Neural-GIF can be trained on raw 3D scans and reconstructs detailed, complex surface geometry and deformations. Moreover, the model can generalize to new poses. We evaluate our method on a variety of characters from different public datasets, featuring diverse clothing styles, and demonstrate significant improvements over baseline methods, both quantitatively and qualitatively. We also extend our model to a multiple-shape setting.

This chapter is based on **Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing, Tiwari *et al.*, ICCV 2021** [TSTPM21]. The code, model, and dataset were made publicly available to stimulate the research. This work was cited more than 100 times. The NeuralGIF-based formulation for animating avatars aligns with and builds upon a growing body of concurrent and subsequent works in neural rendering. Notably, Scanimate [SYMB21] introduced a similar approach for dynamic human modeling, while later methods such as HumanNeRF [WCS*22] and Animatable NeRF [PDW*21] further extended these ideas for animatable NeRF-based human avatars. This convergence highlights the significance and impact of the proposed formulation in the landscape of neural representations for human avatars.

5.1 Introduction

Human avatars enable numerous applications related to augmented and virtual reality, such as telepresence for enhanced communication and entertainment, and have been instrumental to reconstruct and perceive people in images [AMB*19, AMX*18b, BKL*16, HX*20, XZT19, XJS19]. Human shape deforms according to articulation, soft-tissue and non-rigid clothing dynamics, which make realistic animations extremely challenging.

State-of-the-art body models [LMR*15, PCG*19, XJS19] typically learn to deform a *fixed topology template*, usually using linear blend skinning to model articulation, and blendshapes to model non-rigid effects [LMR*15]. This framework has been extended to account for soft-tissue dynamics [PMRMB15] and clothing [MYR*20, PLPM20, SOC19]. However, the major constraint remains unchanged: all deformations must be expressed as offsets from a pre-defined mesh topology. This introduces fundamental limitations when modeling loose or complex clothing, layered outfits, or clothing that exhibits topological variations, such as skirts, coats, or flowing dresses. For example, it would be difficult to model the subjects shown in Figure 5.1, wearing highly complex, deformable, voluminous clothing, using a single, fixed-resolution mesh. Furthermore, every type of deformation , whether it is soft-tissue jiggle, clothing flutter, or fold, requires a different model formulation. Training these fixed-template based models also relies heavily on expensive pre-processing pipelines that bring raw 3D/4D scans into dense correspondence with the template [LMR*15, PMRMB15], a notoriously fragile and error-prone process. This task becomes significantly more difficult when scans involve cloth with large deformations or clothing with different topologies [WHG*24, IRG*23a], where correspondence is not only hard to establish but often ill-defined. As we also discussed in Section 4.7.1, such mesh-based assumptions limit the model's ability to generalize to real-world clothing variability and dynamics, necessitating more flexible and expressive representations. To overcome the limitations of mesh-based models, several works have turned to implicit function rep-

resentations for modeling human shape [SHN*19, SSSJ20, BSTPM20a, CAPM20]. These methods regress continuous occupancy or signed distance fields from input modalities such as images or 3D point clouds, enabling detailed surface reconstruction without requiring template correspondences. These approaches are able to capture high-frequency details such as cloth wrinkles, hair, and subtle clothing structure—features that traditional models struggle with. However, these implicit reconstructions are typically static: they produce a single 3D surface per observation and do not support articulation or animation. As a result, while they overcome some representational limitations, they fall short in applications that require animation, retargeting, or real-time interaction.

Building on these observations, we proposed a new model, called Neural Generalized Implicit Functions (Neural-GIF), to animate people in clothing as a function of body pose, without relying on fixed mesh topologies. Our approach diverges from traditional mesh-based pipelines by utilizing a continuous, implicit representation that accurately captures pose-conditioned deformations in clothing and soft tissue with high fidelity. We demonstrate that Neural-GIF can model complex, voluminous clothing, such as loose jackets, flowing skirts, or layered outfits—as well as nuanced body deformations, at a level of detail and realism not easily achievable with conventional template-based methods. In contrast to most prior work [BTTPM19, MYR*20, TBTPM20], our method does not require registration to a predefined mesh template during training or inference. This is a significant advantage, as the use of mesh templates inherently limits geometric resolution and diversity, often necessitating coarse approximations of detailed clothing geometry. Moreover, the registration process itself—used to bring scans into correspondence with a canonical mesh—is not only computationally expensive but also brittle, especially for clothing with large deformations or topology changes [PMPHB17b]. Despite extensive research into both optimization-based [LIPM19b, BTTPM19, AMB*19] and learning-based registration [MCPM24, FKL*23, CPMAMN22] techniques, existing methods either fail to produce high-quality results or lack robustness in the presence of complex poses and diverse clothing styles. We provide qualitative evidence of these limitations in Figure 5.12, where fixed-template-based methods struggle to faithfully explain the geometry of loose clothing, both in terms of explaining the large deformations of clothing geometry and high-frequency fine details.

As previously mentioned, the introduction of neural implicit 3D representations has enabled a paradigm shift away from predefined template meshes and the expensive, error-prone process of registering raw 3D scans. These methods model geometry as continuous functions (e.g., occupancy or signed distance fields), eliminating the need for dense correspondences and allowing flexible, high-resolution surface modeling. Following this shift, several concurrent and follow-up works explored animatable neural implicits [DLJ*20, SYMB21, MZBT21, MSB*22, CJS*22], further demonstrating the potential of this representation for human modeling. We introduced Neural-GIF, an animatable neural implicit model that captures complex pose-dependent clothing and soft-tissue deformations without relying on any mesh-based registration pipeline. Unlike prior mesh-based approaches for modeling *people in clothing* [MYR*20, PMPHB17a, BTTPM19], our method requires only the pose and shape parameters (e.g., θ, β from SMPL) for each training sample, and no template fitting or correspondence pre-processing. A key strength of Neural-GIF lies in its ability to

represent garments of varying topology—such as coats, skirts, and loose jackets—within a unified framework. This level of topological flexibility is difficult, if not impossible, to achieve with fixed-topology mesh models.

At its core, Neural-GIF learns a signed distance field (SDF) of the posed human surface using a neural network. However, directly mapping pose to surface geometry via SDF regression is a highly underconstrained problem that does not generalize well across complex poses and clothing types, as observed in [DLJ*20]. To address this, we draw inspiration from template-based models that factorize deformations into articulated motion and residual non-rigid effects. Neural-GIF generalizes this idea in the implicit setting by learning a deformation field that maps query points from the posed space into a canonical space, where a pose-dependent deformation field is learned. Since our model queries operate in the posed space, there is no need for data or output canonicalization at any stage of training/inference, and we train the model based solely on posed meshes.

This deformation-based canonicalization significantly eases learning by allowing the network to focus on modeling a more constrained canonical shape, which can then be flexibly articulated and non-rigidly deformed using a learned field. Our model design and naming are inspired by the Generalized Implicit Functions formulation [SP91], which shows that diverse shape deformations can be captured by learning spatial transformations of a canonical implicit function. We elaborate on the conceptual parallels between our volumetric representation and traditional mesh-based body models in Section 5.4.2, where we position Neural-GIF as a natural extension of the SMPL formulation into the implicit domain.

In summary, our contributions are:

- **Neural-GIF**, a novel implicit-based re-posable character model, directly learned from raw 3D scans. Our method is capable of representing complex human and clothing geometries with varying topologies, enabling high-fidelity reconstruction and reposing without relying on template meshes.

- We introduce a **canonical mapping network** that learns a continuous backward skinning field in 3D space. This network maps posed 3D points into a canonical space, without requiring explicit supervision or canonicalization of scans. The proposed backward skinning-based formulation is simple yet effective, enabling the learning of more complex deformations.

- We propose a **displacement field network** that refines point positions in the canonical space before Signed Distance Function (SDF) evaluation. This leads to improved surface detail and enables the modeling of high-frequency deformations, such as wrinkles and folds, resulting in realistic clothed human reconstruction.

- We introduce a **ClothSeq**, a dataset of three subjects in highly complex clothing(open jacket, layered clothing, loose shrugs) and diverse poses. This is the first dataset to feature such complex clothing, serving as a benchmark for assessing the capabilities of implicit-based models in modeling large and dynamic deformations.

We test our method on a variety of scans originating from different datasets such as ClothSeq, CAPE [MYR*20], TailorNet [PLPM20], Dfaust [BRPMB17b], and provide extensive

quantitative and qualitative comparisons. We also extend our formulation to multiple shape settings by adding a shape-dependent displacement field network. Experiments demonstrate that our method generalizes to new poses, models complex clothing, and is significantly more robust and more detailed than existing methods [SYMB21, DLJ*20].

In Section 5.3, we introduce our proposed framework, NeuralGIF, and provide a comprehensive explanation of each of its core components. We also provide training and implementation details in this section. Section 5.4 contextualizes our approach by drawing connections to classical animation techniques, namely skinning-based methods and parametric body modeling. A key concept in our formulation is backward skinning. We review how recent works [CZB*21, SYMB21] have leveraged either backward or forward skinning formulations to achieve either more complex deformation modeling or better pose generalization, and we position NeuralGIF within this evolving landscape of methods. Furthermore, we include a comparative analysis between NeuralGIF and the SMPL body model, highlighting both conceptual and structural parallels and differences.

We also introduce a new dataset featuring highly complex and varied clothing geometries, which is detailed in Section 5.5. In Section 5.6, we present both qualitative and quantitative evaluations of our method, including comparisons with contemporary and prior methods. Additionally, we conduct an ablation study to analyze the contribution and impact of each module within our architecture. While our primary focus is on animating a given subject in fixed clothing, we also demonstrate the applicability of our formulation to separate clothing items. Moreover, we propose an extension to a multi-shape model, where a single model is trained across multiple subjects from the AMASS dataset [MGT*19b]. Finally, in Section 5.7 we conclude by discussing the limitations of our approach, the impact, and the reception of our work since its publication.

5.2 Generalized Implicit Functions(GIF)

Implicit surface representations provide a powerful mechanism for encoding complex geometries using continuous functions. Specifically, an implicit surface is defined by points $\mathbf{p} \in \mathbb{R}^3$, such that the surface is characterized by a scalar occupancy or signed distance function $f(\mathbf{p})$, where $f(\mathbf{p}) = 0$ explicitly describes the surface. Points for which $f(\mathbf{p}) < 0$ lie inside the surface, and points for which $f(\mathbf{p}) > 0$ lie outside.

To extend implicit surfaces to represent shapes undergoing transformations, we generalize the implicit function to arbitrary coordinate frames or posed space. Given a new coordinate frame defined by transformed points $\mathbf{p}' \in \mathbb{R}^3$, we would need to know the occupancy/sdf in this new coordinated frame or posed space, given by $f'(\mathbf{p}')$, where f' is a new function. Designing or learning new f' for all new posed spaces is difficult, if not impossible. However, a simple coordinate transformation can resolve this issue. Specifically, we compute occupancy by inversely transforming these points back to the canonical coordinate frame and evaluating the original occupancy function $f(\mathbf{p})$. We now explain this for simple rigid transformations and displacement field-based transformations.

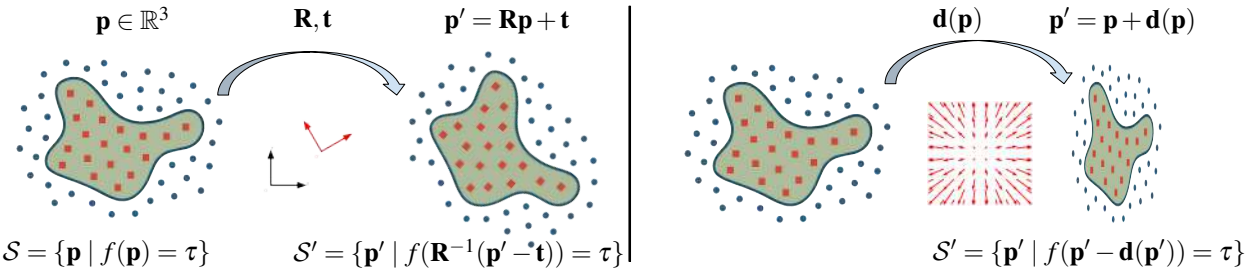


Fig. 5.2 **Generalized implicit function evaluation under different transformations:** Here, we use occupancy as an implicit function. (*Left*) Rigid transformation applies rotation and translation$(\mathbf{R}, \mathbf{t})$ mapping canonical points $\mathbf{p}$ to transformed points $\mathbf{p}'$. Occupancy is evaluated by inverse-transforming points back to the canonical space. (*Right*) Displacement field-based transformation, where a spatially varying displacement field $\mathbf{d}(\mathbf{p})$ maps canonical points to displaced points $\mathbf{p}'$. Occupancy is similarly determined by evaluating the original implicit function after inverse displacement.

5.2.1 GIF for Rigid Transformations

We define our original space(canonical space) by points $\mathbf{p}$, where the implicit surface is given by :

$$\mathcal{S} = \{\mathbf{p} \mid f(\mathbf{p}) = \tau\} \tag{5.1}$$

where τ is the level set. In the case of a signed distance function, this value is usually 0, and for occupancy, this is 0.5.

Now we apply a rigid transformation to our surface, and the points of the transformed surface are given by $\mathbf{p}' = \mathbf{R}\mathbf{p} + \mathbf{t}$, where $\mathbf{R}$ is the rotation and $\mathbf{t}$ is the translation. Following the idea of transforming the query point back to the original canonical space, the occupancy/sdf in the transformed space can be simplified as:

$$\mathcal{S}' = \{\mathbf{p}' \mid f'(\mathbf{p}') = \tau\} \tag{5.2}$$

$$= \{\mathbf{p}' \mid f(\mathbf{R}^{-1}(\mathbf{p}' - \mathbf{t})) = \tau\} \tag{5.3}$$

This equation effectively maps the query points $\mathbf{p}'$ from the transformed frame back into the canonical space using the inverse rigid transformation. As a result, we don't need to learn/model f' for the new posed spaces.

5.2.2 GIF for Displacement Field-Based Transformation

Generalized implicit functions can also represent non-rigid deformations through spatially varying displacement fields. Let $\mathbf{p}'$ denote the deformed point in space, and let the displacement field be defined as $\mathbf{d}(\mathbf{p})$, specifying a per-point offset or displacement field. To evaluate the implicit function in the deformed space, we again map the query point back to

the canonical space:

$$\mathbf{p}' = \mathbf{p} + \mathbf{d}(\mathbf{p}) \quad \Rightarrow \quad \mathbf{p} = \mathbf{p}' - \mathbf{d}(\mathbf{p}') \tag{5.4}$$

Accordingly, the transformed surface is defined as:

$$\mathcal{S}' = \{\mathbf{p}' \mid f'(\mathbf{p}') = \tau\} \tag{5.5}$$
$$= \{\mathbf{p}' \mid f(\mathbf{p}' - \mathbf{d}(\mathbf{p}')) = \tau\} \tag{5.6}$$

As with rigid transformations, this formulation enables the direct evaluation of implicit surfaces using the canonical function f, compensating for the deformation. Thus, there is no need to explicitly model a separate f' for each non-rigidly transformed configuration. We use a combination of rigid transformation (for articulation) and displacement field-based transformation in NeuralGIF to model pose-dependent and detailed deformation in canonical space.

Figure 5.2 provides a visualization of the generalized implicit function evaluation under different transformation types. We have used the occupancy function as an implicit function here for the ease and clarity of visualisation, but a similar concept can be extended to the signed distance function. The left part of the figure shows how occupancy is evaluated by applying the inverse of a rigid transformation, comprising rotation and translation, to map transformed points back to their canonical positions. In contrast, the right part of Figure 5.2 shows evaluation under a displacement field-based transformation, where each point is displaced by a spatially varying vector and occupancy is again computed by inverting this displacement. Together, the figure emphasizes the central role of inverse mapping in this generalized framework. In our model, we use a combination of both, as explained in the following sections.

5.3 Animating People in Clothing

Our method, called Neural-GIF, is a neural model based on generalized implicit functions for animating people in clothing. As mentioned earlier, our method traces its roots back to a seminal paper from 1991 titled "Generalized Implicit Functions" [SP91]. We have discussed this in Section 5.2. In this section, we discuss our method, which extends the Generalized Implicit Function for animating people in clothing. Given a pose(θ) as input, Neural-GIF predicts SDF, whose zero level set represents the character's surface $\mathcal{S}$. Without loss of generality, we use SMPL [LMR*15] for our pose representation. To train Neural-GIF, we require sequences of 3D scans of a subject in fixed clothing and the corresponding SMPL parameters. We do not require non-rigid registration and un-posing as a pre-processing step [PLPM20, PMPHB17b, PMRMB15, LCT18a], which is a tedious task and prone to error. Since learning the mapping $\theta \mapsto \text{SDF}(\mathcal{S})$ is hard (large variations due to articulation, fine non-rigid deformations due to soft-tissue and clothing) and often results in missing limbs, we factorize the motion to learn a *deformable SDF in a canonical* (unposed) space of the character. Three neural functions are composed to obtain the pose-dependent SDF:

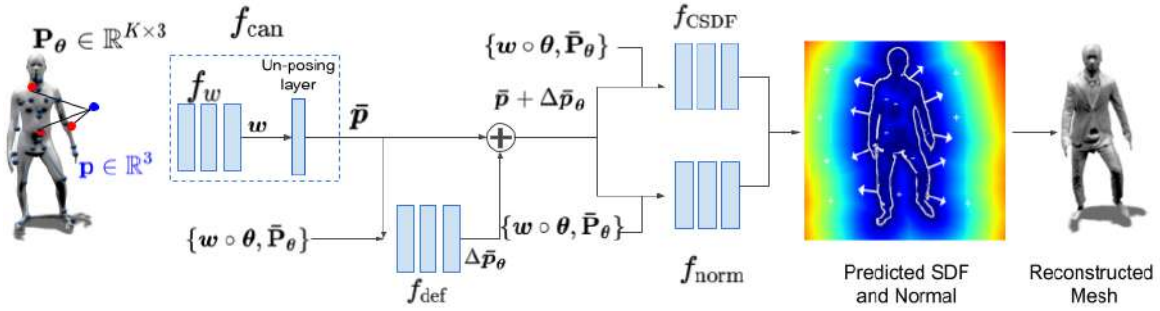


Fig. 5.3 **Neural Generalized Implicit Function:** Neural-GIF first maps every point $\mathbf{p}$ in 3D posed space to the canonical space by learning the associations between the point and the human body with $f_{\mathbf{w}}$ based on a pose encoding ($\mathbf{P}_\theta$) as input. Then we predict a displacement field model (f_{def}) learns a pose-dependent deformation field in canonical space ($\Delta\bar{\mathbf{p}}_\theta$), conditioned on weighted pose parameters ($w \circ \theta$) and pose encoding ($\bar{\mathbf{P}}_\theta$) in the canonical space. We add the displacement to the unposed point before evaluating the SDF using our canonical SDF network (f_{CSDF}). We also employ a normal prediction network (f_{norm}) to obtain results with fine-level details. Both f_{CSDF} and f_{norm} take weighted pose and pose encoding as input.

- **Canonical mapping network**(Section 5.3.1): This network maps every point in the posed 3D space to the canonical unposed space by learning point to human body associations.
- **Displacement field network**(Section 5.3.2): It models non-rigid pose-dependent deformations (soft-tissue, cloth dynamics). Specifically, it predicts a continuous displacement field for points in the canonical space.
- **Canonical SDF**(Section 5.3.3): This model takes the transformed point with the above networks, and a pose encoding as input to predict the desired *signed distance* for every query point. We additionally predict a surface normal field as a function of pose in canonical space to add realism in the results.

5.3.1 Canonical Mapping Network

Our canonical space is characterized by a T-pose of character. Learning in this space enables the network to discover a mean shape that can be deformed and articulated, which facilitates pose generalization. Formally, the main task of the network is to map every point in the posed space (including off-surface points) to the canonical space $\mathbf{p} \mapsto \bar{\mathbf{p}}$. This requires associating the point $\mathbf{p}$ with the body, and inverting the articulated motion. While we can associate every point with its closest SMPL body model point, this is often ambiguous in occluded regions near the armpits or between the legs (see Figure 5.10). Consequently, we propose to learn this *association as a function of pose*.

Instead of simply using pose θ as input, we use a body-aware pose representation. First, we take K joints the SMPL mesh ($\mathbf{j_i}, i \in \{1 \dots K\}$), and then define a pose encoding $\mathbf{P}_\theta \in \mathbb{R}^{K\times 3}$, where $\mathbf{P}_{\theta,i} = ||\mathbf{p} - \mathbf{j_i}||$.

Given a query point in space and its pose encoding, the *canonical mapping network* predicts the blend weights $\mathbf{w} \in \mathbb{R}^K$ of query point $\mathbf{p}$ and transforms it to the canonical space using a differentiable un-posing layer:

$$\mathbf{w} = f_{\mathbf{w}}(\mathbf{p}, \mathbf{P}_\theta) \tag{5.7}$$

$$\bar{\mathbf{p}} = \left(\sum_{i=1}^{K} \mathbf{w}_i \mathbf{B}_i \right)^{-1} \mathbf{p}, \tag{5.8}$$

where $\mathbf{B}_i$ is the transformation matrix for joints $i \in \{1 \dots K\}$, and $\mathbf{w}_i$ its blend-weight. Eq. (5.7) and (5.8) define the canonical mapping network $f_{\text{can}} : (\mathbf{p}, \mathbf{P}_\theta) \mapsto \bar{\mathbf{p}}$.

5.3.2 Displacement Field Network

A surface expressed implicitly as the zero level set $\{\mathbf{x} \in \mathbb{R}^3 \,|\, SDF(\mathbf{x}) = 0\}$ can be deformed to obtain a new surface by applying transformations to the points themselves. By applying a displacement $\mathbf{x}' = \mathbf{x} + \Delta\mathbf{x}$ we will effectively displace the zero level set $\{\mathbf{x} \in \mathbb{R}^3 \,|\, SDF(\mathbf{x}') = 0\}$, and consequently the surface defined. To model subtle non-rigid deformations, we propose to explicitly learn a pose-dependent displacement field in the canonical space:

$$\Delta\bar{\mathbf{p}}_\theta = f_{\text{def}}(\bar{\mathbf{p}}, \mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta) \tag{5.9}$$

where $\mathbf{w} \circ \theta$ is weighted SMPL pose parameter and $\bar{\mathbf{P}}_\theta$ is pose encoding computed in the canonical space. We denote the space of pose encoding as $(\mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta) \in \mathcal{P}$. Note that these are *not per-vertex displacements*, but a *learned continuous displacement field* $\mathbb{R}^3 \mapsto \mathbb{R}^3$, and hence we are not restricted to a fixed mesh with a fixed number of vertices.

5.3.3 Canonical SDF and Normal Prediction

Given points in canonical space and the pose encoding $(\mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta) \in \mathcal{P}$, we learn a signed distance field in canonical space. The canonical SDF ($f_{\text{CSDF}} : \mathbb{R}^3 \times \mathcal{P} \to \mathbb{R}$) is parameterized with a fully-connected neural network. As seen in Fig. 5.3, the network predicts SDF values for points in the canonical space with the following equation:

$$d^* = f_{\text{CSDF}}(\bar{\mathbf{p}} + \Delta\bar{\mathbf{p}}_\theta, \mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta) \tag{5.10}$$

where the canonical point $\bar{\mathbf{p}} = f_{\text{can}}(\mathbf{p}, \mathbf{P}_\theta)$, and the displacement $\Delta\bar{\mathbf{p}} = f_{\text{def}}(\bar{\mathbf{p}}, \mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta)$ are computed with the canonical mapping and displacement field networks respectively. This formulation has several nice properties. First, the network can implicitly learn a canonical shape of clothing/human, without ever explicitly unposing the scans. Note that if the motion consists exclusively of only articulation and displacement, then we could directly learn $f_{\text{CSDF}}(\bar{\mathbf{p}} + \Delta\bar{\mathbf{p}})$ without further conditioning (in practice we condition on pose to give the

network the flexibility to learn shapes which can not be modeled via displacements from a canonical shape). Second, we can supervise Eq. (5.10) with distances computed in the original posed scans.

In addition to signed distances, we predict the normals for points near the surface to generate more realistic-looking results. We predict normals in canonical space, using f_{norm} and then transform it to pose space using blend weights, computed from Eq. (5.7) and $\mathbf{B}_i$:

$$\mathbf{n}^* = \sum_{i=1}^{K} \mathbf{w}_i \mathbf{B}_i f_{\text{norm}}(\bar{\mathbf{p}} + \Delta\bar{\mathbf{p}}_\theta, \mathbf{w} \circ \theta, \bar{\mathbf{P}}_\theta) \tag{5.11}$$

5.3.4 Multi-Subject Model using Neural-GIF

We extend our proposed framework for a multi-subject setting, *ie*we learn multiple shapes in one single model. For this, we draw inspiration from SMPL [LMR*15] model, where per-vertex shape-dependent displacement is introduced, and extend this idea to a continuous shape-dependent displacement field. To incorporate a shape-dependent displacement field, we introduce another displacement network($f_{\text{def}-\beta}$), which predicts a continuous displacement field in $\mathbb{R}^3$, given a shape parameter(β) as input:

$$\Delta\bar{\mathbf{p}}_\beta = f_{\text{def}-\beta}(\bar{\mathbf{p}}, \beta) \tag{5.12}$$

We further modify our model by taking SMPL shape(β) as additional input to each sub-network and adding $\Delta\bar{\mathbf{p}}_\beta$ to $\bar{\mathbf{p}} + \Delta\bar{\mathbf{p}}_\theta$ before applying $f_{\text{CSDF}}(*)$ and $f_{\text{norm}}(*)$.

5.3.5 Training and Implementation Details

In this section, we explain the network architecture, followed by details on training and inference.

Network details: The model consists of four MLP modules, 1) Skinning weight prediction network(f_{w}), displacement prediction network (f_{disp}), canonical SDF prediction network (f_{CSDF}), and a normal prediction network (f_{norm}). Input to each network is first passed to a sinusoidal positional encoding which builds high-frequency features [MST*20] by mapping each coordinate and joint-locations value through multiple sin/cos bands (15 frequencies for f_{w}, 20 for f_{CSDF}, 75 for f_{disp}), while also preserving the raw inputs. f_{w} has 5 linear layers with LeakyReLU activations, and outputs a 24-way softmax distribution of blend weights. Those weights blend the 24-part transforms and invert the SMPL pose to map each point into a canonical T-pose. f_{disp}, shares the same five-layer MLP structure as the weight prediction network and predicts a small 3D displacement on each canonical point to capture non-linear deformation. The displaced points then feed into f_{CSDF}, another five-layer MLP, which regresses a scalar signed distance per point. f_{disp} is similar to f_{CSDF}, except it predicts 3 dimensional normal vector.

Training: Each of the sub-networks is implemented using an MLP. Following [MST*20], we use a positional encoding for query point $\mathbf{p}$. We follow a multi-stage training regime for faster convergence and stable training. First, we pre-train the canonical mapping network (f_{can}) using SMPL meshes generated from the training motion sequences and initially supervise blend weights with ground truth (only known at the SMPL surface). Then we train the canonical mapping network (f_{can}) and the canonical SDF (f_{CSDF}) end-to-end. Then in the final step, we freeze the canonical mapping network (f_{can}) and train (f_{CSDF}, f_{def}) together end-to-end. For training the normal prediction network, we train only for points which are near the surface, *i.e.* $|d^*| < \delta$, once the loss of f_{CSDF} has stabilized. We train our network using the following point-wise losses:

$$\mathcal{L}_{\text{SDF}} = ||d^* - d||_2 \qquad \mathcal{L}_{\text{norm}} = 1 - \mathbf{n}^T \mathbf{n}^* \tag{5.13}$$

where d^*, d represent the predicted and ground truth signed distances respectively, and $\mathbf{n}^*, \mathbf{n}$ denote the predicted and ground truth normals for a point in the posed space. We also added a regularizer for the pose-dependent displacement field network, given by:

$$\mathcal{L}_{\Delta\bar{\mathbf{p}}_\theta} = ||\Delta\bar{\mathbf{p}}_\theta||_2^2 \tag{5.14}$$

We generate training data by sampling points on the surface of the ground-truth mesh and randomly displace these points with noise $\sigma = \{0.01, 0.1\}$. Similar to [CMPM20], the ground-truth signed distances at these samples are computed by casting randomized rays and checking the parity [JKS13]. We use $K = 24$ joints for fully body meshes and follow [PLPM20] for skirt and shirt meshes. We implement each network using MLPs with 3 hidden layers. We use learning rate, $\eta = 1e-4$, batch size of 8.We train for 6-7 hours for ClothSeqand 13-15 hours for other datasets, on NVIDIA Quadro RTX 8000.

Inference: For generating meshes from our method, we first sample points in space and predict d^*, using Eq. (5.10). We reconstruct our meshes using marching cubes [LC87] on the predicted SDF. Our reconstructed meshes are at 256 voxel resolution. Then we predict normals for the vertices on the reconstructed surface, using f_{norm}. Note that, unlike [SYMB21], we directly query our point in pose space, so there is no need for an extra re-posing step after mesh extraction.

5.4 Skinning & Displacement Fields: Bridging Classical and Implicit Models

In this section, we discuss two key components of our model and draw analogies with classical and recent works. First, in Section 5.4.1, we discuss the concept of backward skinning, which is also employed in concurrent works such as [SYMB21]. We compare this approach with classical forward skinning formulations and highlight neural implicit-based methods that utilize forward skinning [CZB*21]. Then, in Section 5.4.2, we contrast the articulated model used in NeuralGIF with the SMPL (or SMPL+D) body model.

5.4.1 Forward Skinning and Backward Skinning

Here, we compare and contrast the skinning method used in NeuralGIF with that of standard body models such as SMPL [LMR*15] and forward skinning-based neural implicit methods(SNARF [CZB*21]). We will briefly describe the formulations of SMPL, SNARF, and NeuralGIF before making comparisons. For a more detailed overview of skinning and body models, please refer to Section 2.1.1 and Section 2.1.2. Note that for SMPL or SMPL+D-based formulation, we use same notations and deformation as SMPL, as it can be easily extended to SMPL+D, by adding a clothing(identity-specific) displacement field in the canonical space(see Section 4.2.1 and Section 2.1.2 for more details). Ultimately, we discuss the benefits of each formulation.

Traditional forward skinning, used in SMPL, explicitly transforms points from a canonical rest pose (typically T-pose) into a posed space. In contrast, NeuralGIF adopts a **backward skinning** approach, where points from the posed space are mapped back to their canonical space. Unlike classical mesh-based methods, which only require skinning weights defined on mesh surfaces, neural implicit methods require skinning defined volumetrically, which we term **volumetric skinning** or **implicit skinning**. Recent neural implicit-based works [CZB*21, CJS*22] have also used a forward skinning-based formulation. However, these methods face challenges due to the multiple-solution problem, wherein different canonical points map to the same location in the deformed space.

Forward skinning in SMPL: As discussed in Section 2.1.2, skinning in SMPL leads to each vertex in canonical space $\mathbf{p}$ being transformed into $\bar{\mathbf{p}}$ using:

$$\mathbf{p} = \sum_{i=1}^{K} w_i \mathbf{B}_i \bar{\mathbf{p}} \tag{5.15}$$

where w_i is an element of the blend weight matrix $\mathcal{W}$ that represents how much the rotation of part i effects the point and $\mathbf{B}_i$ is the bone trasnformation matrix of i joint.

Forward skinning in Neural Implicits: Recent works [CZB*21, CJS*23, WSGT22], use forward skinning in their formulation and address the aforementioned issue using an iterative root-finding algorithm. Specifically, in SNARF, forward skinning defines the canonical correspondence $\bar{\mathbf{p}}$ of $\mathbf{p}$ implicitly as the root of the following equation:

$$\mathbf{d}_{\sigma_w}(\bar{\mathbf{p}}, \mathbf{B}) - \mathbf{p}' = 0, \quad \text{where} \quad \mathbf{d}_{\sigma_w}(\mathbf{p}, \mathbf{B}) = \sum_{i=1}^{K} w_{\sigma_w, i}(\mathbf{p}) \mathbf{B}_i \mathbf{p} \tag{5.16}$$

Here, finding the canonical point corresponding to an observed posed point requires solving this equation iteratively with root-finding methods, which increases computational expense and complexity.

Backward skinning in Neural Implicits: However, in our backward skinning-based formulation, we learn a mapping from points in posed space to points in canonical space, as

shown in Equation (5.7). For the sake of easy readability, we restate the equation.

$$\mathbf{w} = f_{\mathbf{w}}(\mathbf{p}, \mathbf{P}_\theta) \tag{5.17}$$

$$\bar{\mathbf{p}} = \left(\sum_{i=1}^{K} \mathbf{w}_i \mathbf{B}_i \right)^{-1} \mathbf{p}, \tag{5.18}$$

In Equation (5.15), Equation (5.16) and Equation (5.8), $\mathbf{B}_i$ is the transformation matrix for joints $i \in \{1 \dots K\}$, and $\mathbf{w}_i$ its blend-weight. In case of Equation (5.15), $\mathbf{w}_i$ is defined on the canonical space and doesn't depend on pose. For Equation (5.16), w_{σ_w} is neural blend skinning or skinning weights predicted from networks and only depend on the canonical point location, $\bar{\mathbf{p}}$. In case of Equation (5.18), $\mathbf{w}_i$ is also predicted by the network and defined in the pose space, and hence it depends on the pose. Note that in different poses, the same point in the deformed space will have different skinning weights, because the deformed space itself changes with the pose. We define this using Equation (5.17), which we learn in our proposed formulation. In contrast to forward skinning-based formulations in neural implicits, the backward skinning-based formulation is straightforward and avoids reliance on numerical methods. It tends to capture more dynamic and detailed deformation. We provide a comparison between three paradigms of skinning in Table 5.1.

Why Backward Skinning? As highlighted in our work [TSTPM21] and concurrent studies [SYMB21, CZB*21], backward skinning provides several advantages:

- Flexibility to model more complex and loose clothing.
- More dynamic and detailed deformation.
- Simpler formulation without iterative numerical methods.

However, backward skinning may exhibit limited generalization to unseen poses. We mitigate this by pre-training the skinning weight prediction network to produce predictions that are close to SMPL skinning weights, thereby significantly improving generalization.

Training and Inference speed: Follow-up work, Fast-SNARF [CJS*23], provides an extensive comparison of training and inference times. As shown in this work, the backward-skinning-based method requires a training time of 31 *minutes*, as opposed to 8*hours* for the forward-skinning root-finding–based method. Similarly, for inference, the backward-skinning-based method takes 40.60 ms, compared to 994.01 ms for the forward-skinning root–finding–based method. We highlight that Fast-SNARF introduces several algorithmic and implementation improvements over SNARF, yielding a 150 speed-up and making it comparable to backward-skinning-based methods.

Skinning	Mesh-based Forward [MYR*20, PLPM20]	Implicit Forward [CZB*21, WSGT22]	Implicit Backward [TSTPM21, WCS*22]
Formulation	Direct	Iterative root-finding	Direct inverse mapping
Volumetric Skinning	✗	✓	✓
Pose Generalization	High	High	Moderate (enhanced by pre-training)
Dynamic Deformations	Low	Moderate	High
Clothing Flexibility	Low	Moderate	High
Computational Cost	Low	High	Moderate

Table 5.1 **Comparison of skinning approaches used in *human in clothing* models:** While mesh-based LBS offers simple and low computational cost, implicit methods can model complex and highly dynamic deformation, with the backward skinning-based formulation excelling in capturing dynamics and clothing deformation.

5.4.2 Displacement Field and Blend Shapes

In this section, we discuss how the displacement field network in NeuralGIF, which learns a deformation field in canonical space, relates conceptually to the pose blend shapes of the SMPL body model. While we focus primarily on the pose-conditioned displacement field and pose blend shapes, this formulation can easily be extended to shape-conditioned displacement fields (see Section 5.3.4), analogous to SMPL's shape blend shapes, enabling the modeling of multiple shapes using a single model.

SMPL's pose blend shapes explicitly add deformation to canonical vertices to correct pose-induced deformations, hence acting as additive corrections. This correction is performed in the canonical space, before applying the articulation.

$$\bar{\mathbf{p}}_{\text{corrected}} = \bar{\mathbf{p}}_{\text{canonical}} + B_p(\theta, *), \tag{5.19}$$

where $B_p(\theta, *)$ represents the additive pose-dependent blend shapes and $*$ represent other model parameters. For more details, please refer to Section 2.1.2.

NeuralGIF's implicit displacement field, in contrast, performs corrections implicitly after the backward skinning (canonical mapping). Since the displacement field is learned as a function of canonical points, however, the overall sdf is known for the posed points. Hence, it has an inverse effect. Specifically, the posed shape is first mapped back to the canonical space, after which corrective details are added:

$$\bar{\mathbf{p}}_{\text{corrected}} = \bar{\mathbf{p}}_{\text{canonical}} + f_{\text{disp}}(\bar{\mathbf{p}}_{\text{canonical}}, \theta), \tag{5.20}$$

where represents the learned implicit displacement field conditioned on pose. We compare the pose correctives of SMPL with NeuralGIF in Table 5.2.

The canonical mapping network (backward skinning) and the displacement field network in NeuralGIF significantly enhance flexibility and realism compared to SMPL. NeuralGIF effectively handles arbitrary shapes, complex clothing, and detailed deformation, overcoming the limitations of explicit, mesh-based linear blend shape corrections. SMPL uses linear blend skinning (LBS), which, although efficient, restricts the representation of complex, non-rigid

	Pose Blend Shapes (SMPL(+D)	**Displacement Field (NeuralGIF)**
Representation	Vertex-based explicit blend shapes	Volumetric implicit field
Topological Flexibility	Low	High
Fine-scale Details	Limited	High
Posing Artefacts	High(for complex clothing)	Low(limited by generalization capabilties)
Computational Cost	Low (precomputed blend shapes)	Higher (implicit queries)

Table 5.2 **Pose blend shapes and neural displacement fields for modeling pose-dependent deformations:** While SMPL+D uses explicit vertex-based blend shapes with low computational cost, NeuralGIF leverages an implicit volumetric field, offering higher topological flexibility and fine-detail modeling.

deformations, particularly detailed clothing, without requiring additional template-specific modifications. Thus, NeuralGIF emerges as a powerful alternative to SMPL, particularly in scenarios that require intricate geometries, flexible topology, and realistic animation of clothing and soft-tissue deformation. We provide a summarized comparison between the SMPL-based model and NeuralGIF regarding skinning methods, deformation modeling, training data requirements, and general limitations in Table 5.3.

Unlike SMPL+D based models for modeling *people in clothing*, Neural-GIF:

- Learns continuous deformation fields and canonical mappings from raw scans without requiring predefined templates.
- Handles complex geometries and varying topologies naturally.
- Generalizes pose-dependent soft-tissue and clothing deformations efficiently.

	SMPL+D based model [MYR*20, BTTPM19, AMB*19]	**Neural-GIF** [TSTPM21]
Representation	Mesh-based (explicit)	Implicit (continuous SDF)
Skinning	Linear Blend Skinning	Learned volumetric skinning
Deformation Method	Explicit additive blend shapes	Implicit displacement field
Training Data	Registered Meshes	Raw Scans
Topology	Limited, predefined topology	Arbitrary and flexible
Fine-scale Detail	Limited	High
Computational Complexity	Low (explicit operations)	High (implicit network queries)

Table 5.3 **SMPL+D and Neural-GIF models for clothed human representation:** SMPL+D uses explicit mesh-based deformations with limited flexibility, while Neural-GIF employs implicit volumetric fields, enabling flexible topology, high-detail modeling, and direct training from raw scans, albeit at a higher computational cost.

Dataset	Clothing Type	Complexity	#Subjects	#Frames
Dyna [PMRMB15]	Minimal-clothing	✗ Simple	10	∼40K total
BUFF [ZPBPM17a]	T-shirt, pants, shorts	✗ Simple	6	∼30 per subject
CAPE [MYR*20]	T-shirt, pants, shorts, coat	✗ Moderate	15	∼4000 per subject
ClothSeq (Ours)	Shrugs, puffer, open jackets	✓ High	3	500–1000 per sequence
4D-DRESS [WHG*24]	Dresses, shirt, pants, loose-open shirts	✓ High	32	78K total
CustomHumans [HILJO23]	Shirt, pants, t-shirt, shorts	✓ Moderate	81	∼600
4DHumanOutfit [ABB*23]	Shirt, pants, t-shirt, shorts, dress	✓ High	20	11 motions per outfit
ActorsHQ [IRG*23a]	Causual tops, pants, dresses	✓ High	8	∼11K total

Table 5.4 **Comparison of ClothSeq with other 4D human scan datasets:** Our dataset offers real scans of complex clothing across multiple subjects and temporal sequences, presenting a more realistic and challenging benchmark for dynamic clothing modeling. Datasets released prior to ClothSeq are shaded in grey. Some new datasets like 4D-DRESS [WHG*24], ActorsHQ [IRG*23a] have very loose clothing.

5.5 ClothSeq Dataset

To evaluate our method on complex clothing beyond traditional template-based models, we introduce ClothSeq, a new dataset comprising 3 subjects wearing challenging garments such as shrugs, puffer jackets, and multilayered clothing with a loose-open jacket. Each subject performs motion sequences such as standing poses, arm lifts, and torso twists, captured at high temporal resolution across 500–1000 frames per sequence. We show some frames of all three subjects of ClothSeq in Figure 5.4.

Compared to prior 4D human datasets or human in clothing deformation datasets, ClothSeq stands out by capturing real clothing with high geometric complexity, including thick, non-body-conforming garments with layered structures and self-collisions. We compare ClothSeq with other existing datasets in Table 5.4. BUFF focuses on tight-fitting clothing and body shape recovery, lacking the diversity and structure of real-world garments. CAPE introduces looser clothing and more varied poses, such as a coat, but remains limited to body-conforming outfits. In contrast, ClothSeq uniquely provides high-resolution, temporally coherent scans of real subjects in complex, layered garments like shrugs and puffer jackets—making it a more realistic and challenging benchmark for modeling detailed clothing dynamics beyond the constraints of template-based methods. ClothSeq dataset emphasizes pose-dependent deformation, providing realistic motion sequences that challenge template-based or parametric approaches. Unlike datasets that rely on simulated cloth or simplified clothing geometry, ClothSeq focuses on realistic scans, enabling a more demanding and relevant benchmark for clothing modeling research. We show the SMPL+D registration for a scan from ClothSeq in Figure 5.12, which illustrates that existing template-based registration methods struggle to capture such complex geometry, particularly in the presence of thick, layered garments and significant non-body-conforming deformations.

Fig. 5.4 **ClothSeq dataset**: We propose a dataset which consists of subjects in very complicated clothing, such as layered and puffer jackets and open loose jackets in diverse poses.

More recently, complex and high-quality of 4D human scan datasets were published. 4D-DRESS [WHG*24] presents a large-scale collection of motion sequences with diverse loose garments, such as dresses and open shirts. CustomHumans [HILJO23] captures everyday clothing across a wide subject pool, though with more moderate garment complexity. This consists of body-conforming clothing items only. 4DHumanOutfit [ABB*23] introduces a multi-subject benchmark with repeated outfits across motions. Although this dataset mainly includes conforming clothing items, some subjects are also wearing loose clothing items like dresses. ActorsHQ [IRG*23a] focuses on high-fidelity reconstructions of actors in diverse apparel and performances, providing fine-grained geometry, especially for casual clothing. While these datasets enhance the landscape of clothing capture with increased subject and outfit diversity, ClothSeq still is one of the datasets providing highly complex clothing, offering a unique challenge for dynamic clothing reconstruction and modeling.

Data processing: We use SMPL [LMR*15] based pose representation in our method, and we also need underlying body shape for body-aware pose encoding(Sec. 3.1), so we obtain SMPL (θ, β) parameters for scans. For the CAPE and TailorNet datasets, the SMPL parameters are already provided by the dataset. For others, we fit SMPL [BSTPM20a] to obtain the pose. Specifically, we first optimise the SMPL shape parameter (β) for the first 5 frames to get a rough body shape underneath. We use this in our body-aware encoding of pose. Then we track the pose throughout the sequence following the loss functions used in [BSTPM20a]. Hence, our pose estimation is temporally coherent. Unlike prior work [LCT18a, PLPM20], perfect registrations or predefined templates are not required in our approach.

Training and Evaluation Dataset: For evaluation, we construct a test set by selecting 80–100 diverse frames per subject. To examine generalization, we adopt two different test frame selection strategies: (1) including test frames with temporally nearby frames in the training set, and (2) excluding all frames within ±10 of each test frame from training. This allows us to probe how well the model extrapolates to both seen and unseen pose neighborhoods.

For creating the ground truth training dataset, we randomly sample 100000 points on the surface of mesh and displace these points with noise $\sigma = \{0.01, 0.1\}$. Our final training sample consists of points on the surface, near-surface points($\sigma = 0.01$), and far-away points($\sigma = 0.1$). During each training step, we sample 1024 points from each of these 3 regions. For training a normal prediction network, we only select points near the surface and on the surface of the ground-truth mesh.

5.6 Experiments and Results

Our proposed method can be used to learn non-rigid pose-dependent deformations in human/clothing meshes. We show results of our method on various datasets in Sec. 5.6.1 and compare with [DLJ*20, SYMB21] in Sec. 5.6.3. We also evaluate our model in multiple

shape setting in Sec. 5.6.4.

Full Body Datasets: For evaluation of single-shape and clothing model, we use CAPE dataset [MYR*20, PMPHB17b], DFAUST [BRPMB17b, PMRMB15], TailorNet [PLPM20], and a few human scan sequences with complex clothing that we captured in our stage, which we will refer to as ClothSeq.

For multiple subject setup, we evaluate our method on three datasets: 1) MoVi [GMT*21, MGT*19b], 2) SMPL pose and shape dataset, and 3) DFAUST registrations [BRPMB17b, PMRMB15] and compare with [MZBT21]. For the SMPL pose and shape dataset, we utilize the 890 poses provided in the SMPL model and create 9 different body shapes by modifying the first 3 principal components of the shape space. For the CAPE and TailorNet, the SMPL parameters are available in the dataset. For others, we fit SMPL [BSTPM20a, BSTPM20b] to obtain the pose and shape. We split the dataset such that the test set contains 2 unseen motion sequences for CAPE and DFAUST. For SMPL pose-shape dataset, we follow the train/test split used in [PLPM20]. For MoVi, we follow the train/test split used in [MZBT21]. To demonstrate that Neural-GIF can model complex geometries, loose clothing, and topologies, we test our method on the ClothSeq dataset, which comprises complex clothing items such as jackets and loose shrugs. For DFAUST and CAPE, we evaluate our model on multiple sequences. Specifically, for DFAUST, we evaluate Neural-GIF on 5 motion sequences, by training it 5 times with a different train-test split of 13-1 each time. Similarly, for CAPE, we evaluate Neural-GIF on 6 motion sequences, by training it 6 times with a different train-test split of 34-2 each time.

Clothing Datasets: For individual clothing items, we utilize the TailorNet dataset, which provides a rich set of garment deformations corresponding to various body poses and shapes. To ensure consistency and comparability with prior work, we adopt the same training and testing splits as those used in [PLPM20] for the TailorNet dataset.

5.6.1 Reanimating People in Complex Clothing

In this section we discuss and show results of NeuralGIF for modeling people in complex clothing such as jacket, coat, shurgs *etc*.

Modeling People in Complex Clothing: We use our Neural-GIF to animate characters in new poses. We generate results for various characters of the ClothSeq and CAPE datasets. We provide results on poses of the test set (volleyball and walking sequences) of the CAPE dataset in Figure 5.6 and on the ClothSeq dataset in Figure 5.5. Our method generates accurate reconstructions and generalizes to unseen poses, while preserving the structure of identity and generating fine details. Since we directly learn from scans, which are noisy, *e.g.*, near the hands, feet, and inner thigh, the reconstruction around these areas tends to be a bit noisy and unstructured. Modeling such types of clothing using template-based methods is not straightforward due to the open-layered structure of the clothes on top of the body.

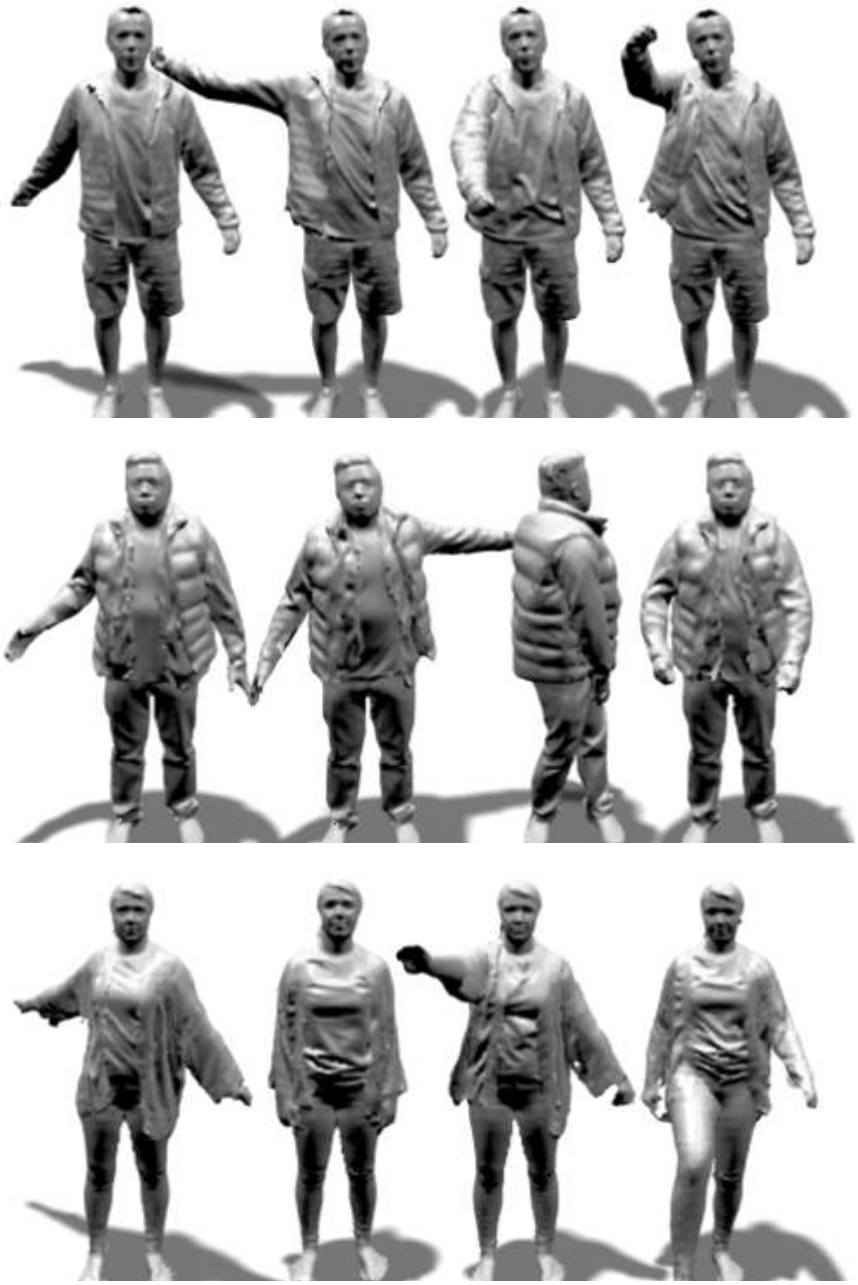


Fig. 5.5 **Animating people in complex clothes:** Results of Neural-GIF for subjects from the ClothSeq dataset. Neural-GIF can model well complex cloth geometries well, like jacket with hoods (*top*), puffer jacket (*middle*) and loose clothes like shrugs (*bottom*).

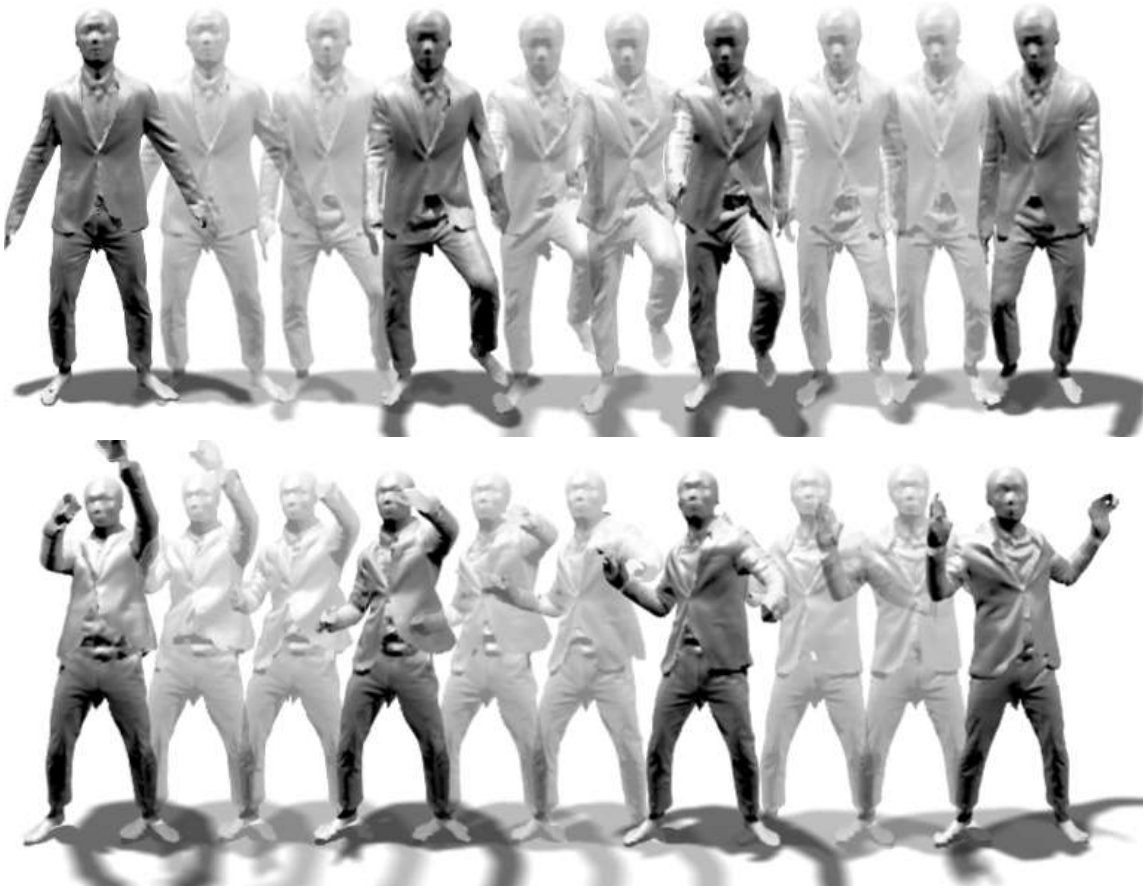


Fig. 5.6 **Reanimating CAPE subjects:** Reposing a character from CAPE dataset [MYR*20] using our proposed method. We show results on the walking motion sequence (*top*) and on the playing volleyball motion sequence (*bottom*).

As a result, a single template cannot explain two extreme deformations of a jacket, (*ie*, on one end when a layer of the jacket is close to the body surface and on the other side of the spectrum when the layer is far from the body surface due to some pose like lifting hands in the air). Our method produces large deformations for loose garments like open jackets and shrugs(Figure 5.5) as opposed to template-based methods. This is possible because we do not rely on blend weights associated with the underlying body, but instead, we predict blend weights for every point in the space as a function of pose, for a given subject. So our network learns to associate each point correctly to joints *e.g.*, a point on the surface of the jacket near the torso will have a different association (weight) in the rest pose and a slightly different association when the person is lifting hands. Our canonical mapping network learns this pose-dependent association from data.

Modeling Separate Clothing Items: Since our method is not limited by a pre-defined template or topology, we can also train Neural-GIF on the clothing meshes of shirts and skirts from the TailorNet dataset, and for completeness, also compare to TailorNet [PLPM20]. It is essential to note that such a comparison is not entirely fair to our method. First, the training meshes are of a fixed resolution dictated by the template used in TailorNet, so our model is upper-bounded by this. Second, registration is noise-free because meshes are obtained from physics simulations (an ideal case for template-based methods), which allows for the mapping of every shape to a vector, making learning significantly easier for TailorNet. Correspondence is known for TailorNet, whereas it is not known for Neural-GIF. Despite this, qualitatively Neural-GIF is comparable to TailorNet [PLPM20] as seen in Figure 5.7. We can see that our method generalizes to new unseen poses and predicts fine details and deformations. Quantitatively, we find that point to surface distance for TailorNet results is slightly better than our method. We produce 11.14 mm and 10.98 mm error for skirt and shirt clothing, while TailorNet's results are 9.82 mm and 7.28 mm respectively. Another source of slight inaccuracy originates from our mesh reconstruction method for open surfaces. Since shirts/skirts are a thin layered surface, we reconstruct the mesh using a small threshold of 0.01mm in marching cubes, which might result in slightly thicker surfaces, and could be addressed by switching to unsigned neural distance fields [CMPM20]. This comparison shows that when registration is possible (easy clothing animated with physics simulation) it helps learning.

5.6.2 Modeling Soft-tissue Dynamics

Neural-GIF models high-frequency pose-dependent deformations for various shapes like full clothing, jackets, and skirts. We now evaluate how well this representation can model soft-tissue dynamics, which are highly non-linear and high-order pose-dependent deformations. Thus, we evaluate our method on this task using the DFAUST [BRPMB17a] dataset. We show our results on DFAUST [BRPMB17a] in Figure 5.8. We evaluate our method on a single subject with 14 sequences, which demonstrate extreme soft-tissue dynamics. In Figure 5.8, we provide results for 3 sequences and observe that the network predicts soft-tissue dynamics/deformations. Our normal prediction network also helps in producing more

Dataset \ Model	NASA [DLJ*20]			SCANimate [SYMB21]			**Ours (Neural-GIF)**		
	Point2Surface ↓	IoU ↑	F-Score ↑	Point2Surface ↓	IoU ↑	F-Score ↑	Point2Surface ↓	IoU ↑	F-Score ↑
CAPE [MYR*20]	10.67	0.918	94.32	**5.82**	**0.957**	**98.51**	**5.86**	**0.957**	**98.53**
ClothSeq	23.26	0.780	57.29	7.32	0.953	97.32	**4.73**	**0.967**	**99.15**
DFAUST [BRPMB17a]	10.52	0.939	95.48	3.79	0.971	99.50	**3.21**	**0.972**	**99.56**

Table 5.5 **Comparison of NeuralGIF with prior and concurrent works:** We quantitatively compare the results of our method with NASA [DLJ*20] and SCANimate [SYMB21]. We report point to surface distance (in mm) and IoU and F-Scores(%) for comparison.

realistic renders. We notice that since deformation in the facial region is not controlled by any keypoint, and also the subjects' facial expressions change over time, we tend to produce smooth results around the face.

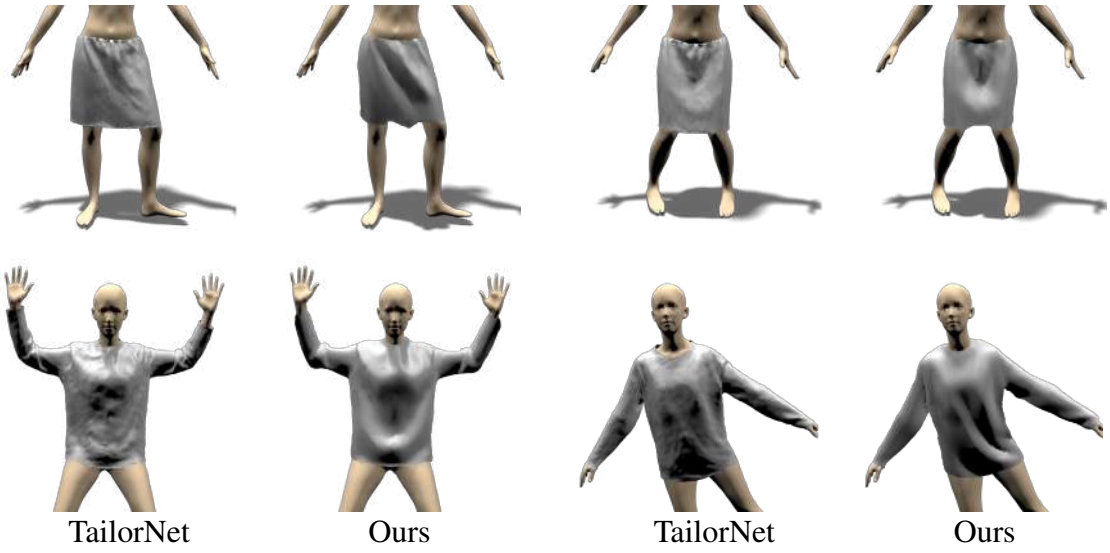


Fig. 5.7 **Comparison with TailorNet [PLPM20] on separate clothing items:** Qualitative comparison on skirt (*top*) and shirt (*bottom*) meshes from the TailorNet dataset. TailorNet produces unsmooth surfaces with unnatural wrinkles, while NeuralGIF generates realistic and high-quality deformations.

5.6.3 Comparisons and Ablation Studies

In this section, we first present a comprehensive comparison of our method with state-of-the-art approaches from both prior and concurrent works, highlighting the improvements and distinctions in terms of realism, and generalization. Following this, we conduct a series of ablation studies to analyze the contribution of each component of our model, thereby gaining deeper insight into the effectiveness and robustness of our overall design.

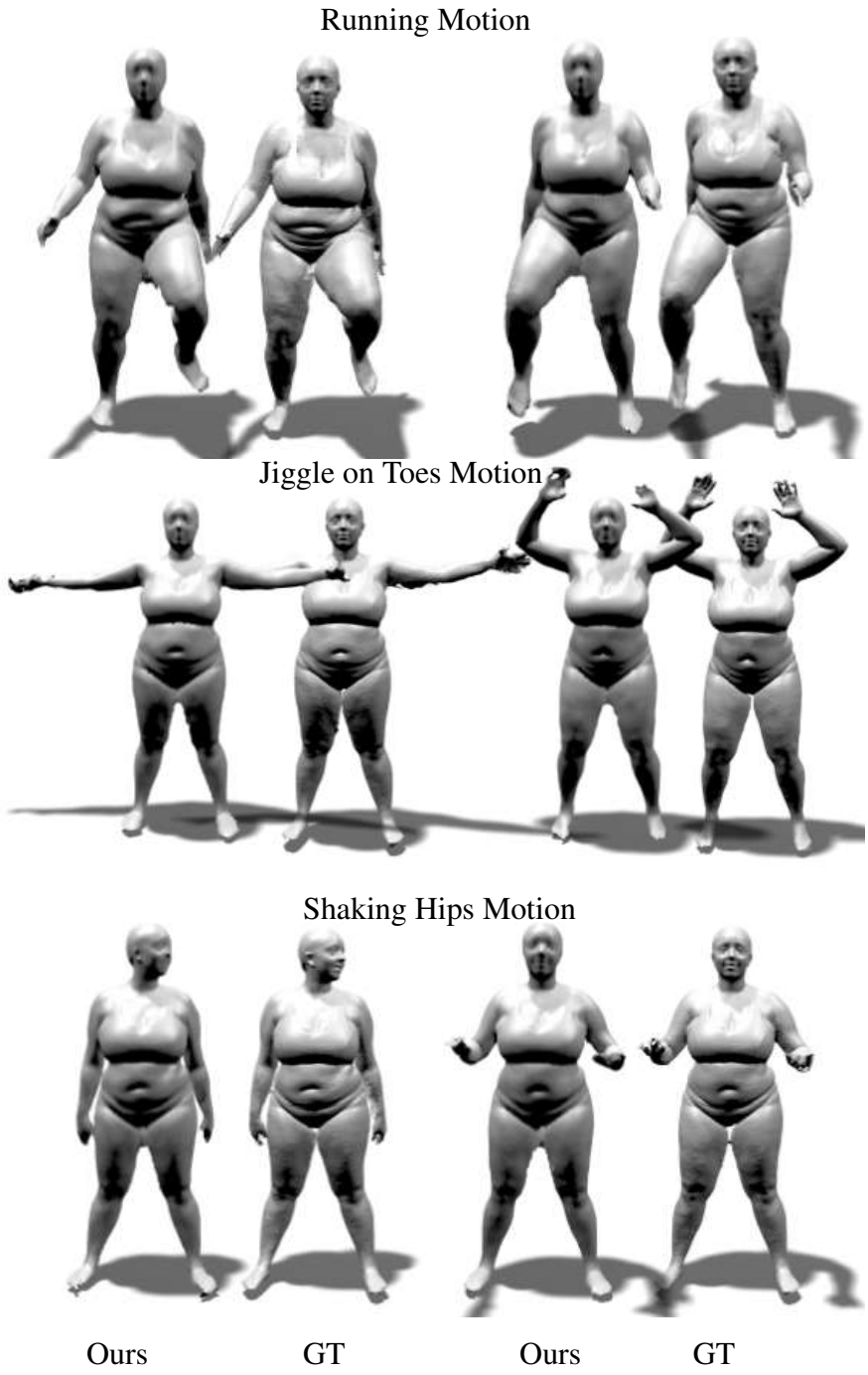


Fig. 5.8 **Results on DFAUST [BRPMB17a]:** We show our results on three different motion sequences (running, jiggle on toes, shake hips) from DFAUST and compare with ground-truth DFAUST data and visualize results in (prediction, GT) pairs. Neural-GIF is able to retain pose-dependent soft-tissue deformations present in the ground truth scans.

Comparison with SoTA: For the single shape model, we compare our method with [DLJ*20, SYMB21] and provide the quantitative evaluations in Table 5.5. In NASA [DLJ*20], pose-dependent occupancy is represented using part-based implicit functions. NASA produces good results for human meshes from [BRPMB17a], however, it cannot model fine details and complex or loose clothing like jackets. We observe from Figure 5.11 that NASA generates rigid and overly smooth results, compared to our model. NASA also suffers from part-based artifacts, such as intersecting body parts in poses like bending elbows or knees. This problem is even more prominent if there are significantly fewer data points for such poses. Since we learn the whole body shape using a single network, our approach does not encounter such issues. From Table 5.5 we observe a significant drop in terms of IoU on ClothSeq when using NASA, which is because the network cannot learn the loose deformations of jackets and hoodies, as one can observe in Figure 5.11. Quantitatively, SCANimate [SYMB21] performs better than our method on CAPE, as seen in Table 5.5, first row, but we notice that results of SCANimate are less detailed and have posing artifacts due to LBS, as highlighted in Figure 5.9. We also show significant improvements in terms of the quality of details produced by our method. We attribute this to the combination of the displacement field network, the normal prediction, and positional encoding of the query point.

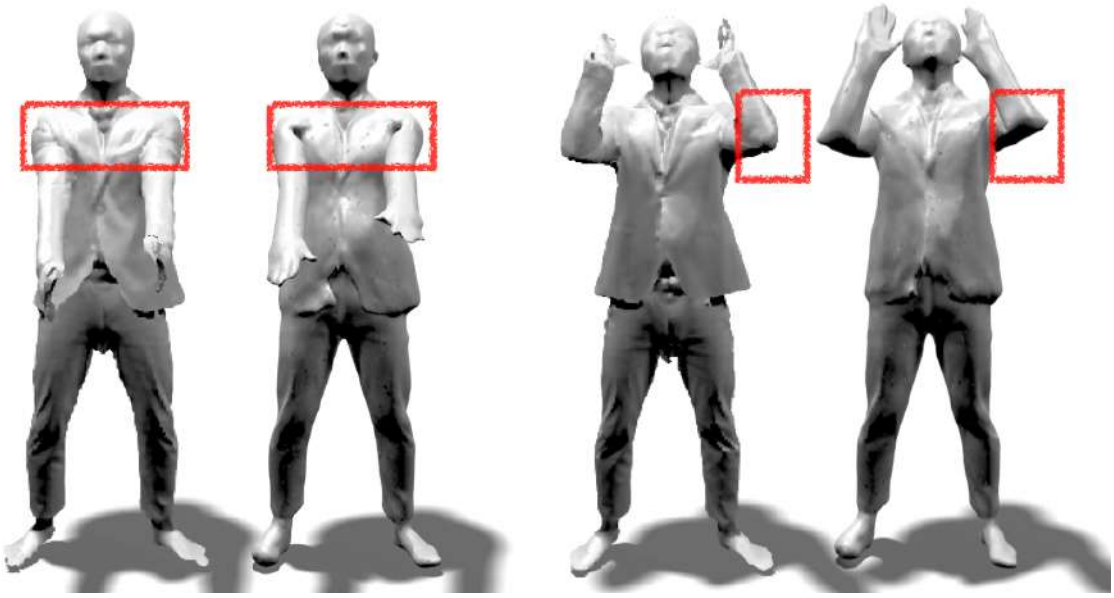


Fig. 5.9 **Comparison with SCANimate:** We compare the results of our method on the CAPE dataset with SCANimate [SYMB21]. Our model preserves more details and does not have posing artifacts.

Ablation studies: We now evaluate the significance of each module in our proposed method and report the result of ablation in Table 6.6 and our complete model in Table 5.5(last column). First, we simply use the Canonical SDF to learn the distance field in the canonical space. We find the blend weights for every point in space ($\mathbf{p}$), by associating the blend weight

Dataset \ Model	SMPL weights		Canonical mapping network	
	Point2Surface ↓	IoU ↑	Point2Surface ↓	IoU ↑
CAPE [MYR*20]	12.93	0.866	**7.93**	**0.955**
ClothSeq	11.89	0.874	**9.18**	**0.963**
DFAUST [BRPMB17a]	18.38	0.806	**3.28**	**0.972**

Table 5.6 **Ablation on learned v/s diffused skinning weights:** Quantitative comparison between using nearest-neighbor-based SMPL weights and our learned canonical mapping network for predicting skinning weights. Our learned skinning weights consistently outperform the baseline, significantly improving reconstruction accuracy and surface alignment.

of nearest point on the SMPL body surface. We transform $\mathbf{p}$ to $\bar{\mathbf{p}}$ using Eq. (5.7) and calculate the SDF using Eq. (5.10). In Figure 5.10(left), we observe that if we simply use nearest point-based blend weights, we obtain wrong reconstructions (space between the legs). We observe that for this case (Table 6.6, col.1), the IoU is relatively low as compared to the other two versions of our model and also lower than NASA(see Table 5.5) on DFAUST. This is because of incorrect body part associations, as shown in Figure 5.10 and explained above. This problem is most prominent near the inner thighs and especially when both legs are not performing a symmetrical motion. The test split in DFAUST consists of jumping, running, etc, the points between the legs randomly change their association between left and right legs and are mapped to a point in the canonical space, which may lie inside the canonical shape, for a given pose.

We alleviate this problem with our canonical mapping network to predict these blend weights (see Table 6.6, col.2). The network learns correct associations based on data. Learning the SDF in the canonical space simplifies the learning process, as the network learns a strong prior for the shape of the subject. We observe from Figure 5.10(middle), that just using Canonical SDF + Canonical mapping network, does not produce fine pose-dependent deformations. Hence, we introduce a displacement field network to learn a pose-dependent displacement field in the canonical space for small non-rigid deformations and a normal prediction model. We observe that having a separate model to incorporate such fine details and training the network in a multi-step regime (Section 5.3.5), helps to learn these deformations.

Comparison with NASA: We use the publicly available code of NASA [DLJ*20] and train the deformable NASA model to compare with our method. We use all the default hyperparameters. For DFAUST [BRPMB17a], we use our data generation step to create training samples.

Comparison with Dyna: We also compare our results with Dyna [PMRMB15], which is a template-based model for representing soft-tissue dynamics using a second-order autoregressive model. Since the code is not publicly available, we compare with the results available in Dyna video/presentations. We observe that our method can produce significantly higher details and finer results compared to Dyna, and the results appear more realistic.



Fig. 5.10 **Ablation on learned v/s diffused skinning weights:** Canonical SDF+SMPL weights result in wrong association and hence result in wrong occupancy (*left*). Our canonical mapping network solves this problem by learning blend weights from data but produces smooth results (*middle*). Finally, the displacement field network and normal prediction network help in generating small pose-dependent details (*right*).

However, since we directly learn from scans, we tend to produce results with noisy hands and feet.

Results on TailorNet meshes: We also show pose-generalisation of our method for the shirt and skirt dataset and compare it with TailorNet results in Figure 5.7. Note that we do not perform any post-processing steps to remove the intersection between the body and garment.

Limitations of SMPL+D methods for Capturing Complex Clothing

In Figure 5.12, we compare a raw scan with its SMPL+D registration and the result of our NeuralGIF model. This comparison highlights two fundamental limitations of relying on SMPL+D registration as ground truth for learning-based models. First, although SMPL+D is widely used as a pseudo ground truth in garment learning frameworks, it inherently struggles to represent complex garment geometry, particularly for non-body-conforming clothing, such as open jackets. As the jacket detaches further from the underlying body surface, the mesh-based deformation model fails to capture the resulting folds, thickness variations, and volumetric flow of the fabric, which are critical for realism. Second, SMPL+D uses a fixed-resolution mesh, which restricts its ability to model fine-grained surface details, such as wrinkles, seams, or layered textures. Consequently, any method supervised on such ground truth inherits these limitations, and hence learning models cannot surpass the expressiveness or fidelity of their supervision data. Our results demonstrate that while the NeuralGIF

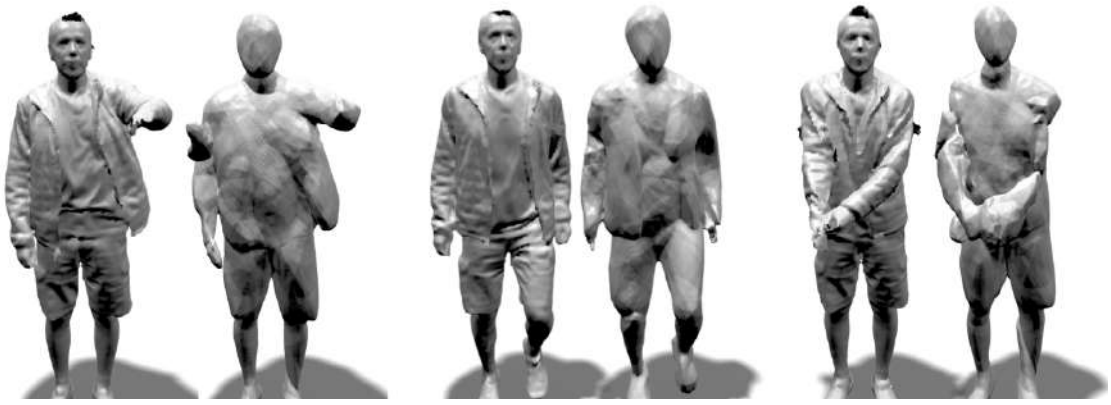


Fig. 5.11 **Comparison with part-based model (NASA [DLJ*20]):** We compare the results of our method on the ClothSeq dataset with NASA [DLJ*20]. Our model reconstructs complex clothing without any part-based artifacts and generates realistic pose-dependent deformations.

Model / Dataset	LEAP [MZBT21]		Ours (Neural-GIF)	
	Point2Surface ↓	IoU ↑	Point2Surface ↓	IoU ↑
DFAUST [BRPMB17a]	3.42	0.958	**3.35**	**0.963**
MoVi [GMT*21]	**3.19**	**0.969**	3.20	**0.969**
SMPL	3.26	0.968	**3.18**	**0.971**

Table 5.7 **Comparison with multi-shape model (LEAP [MZBT21]):** Quantitative results on three different datasets(DFAUST, MoVi, and SMPL).Neural-GIF achieves competitive or superior performance, demonstrating its effectiveness in capturing pose-dependent surface deformations.

model achieves more nuanced representations than the registration, its potential remains fundamentally constrained by the quality and expressivity of the provided mesh supervision.

5.6.4 Multi-Subject Model

As mentioned in section 5.3.4, we extend our formulation to incorporate multiple body shapes in one model. We compare with the concurrent method LEAP [MZBT21] and show the quantitative evaluation in Table 5.7. We model body shape variation as a continuous displacement field in canonical space. Even with a simple formulation, we perform comparably or better than LEAP. We also observe that our training is more stable than LEAP.



Fig. 5.12 **Comparison with mesh based method:** Comparison of the raw scan (*left*), SMPL+D registration (*middle*), and our NeuralGIF result (*right*). The registration fails to capture the complex geometry of the open jacket and fine details due to mesh resolution and body-conforming nature of SMPL+D, which also limits the expressivity of models trained on it.

5.7 Conclusion

In this chapter, we presented Neural-GIF, a novel implicit function-based approach for animating humans in complex clothing driven by body pose. Neural-GIF can be trained end-to-end and learned directly from scans without the need for a complex registration process. The core of our method involves mapping posed space to a canonical space via a backward skinning formulation, followed by predicting pose-dependent displacement fields to capture intricate non-rigid clothing and soft-tissue deformations. Neural-GIF defines shape as displacement and articulation deformations of a canonical shape, a concept that has been widely used for parametric meshes, which we generalize here to neural implicit function learning. We provide an extensive discussion around this in Section 5.4. This approach facilitates the animation of detailed garments, including loose, layered, and highly deformable clothing, at a fidelity difficult to achieve with traditional template-based approaches.

Neural-GIF is validated on diverse datasets, including ClothSeq, CAPE [MYR*20], and DFAUST [BRPMB17a], and shows improved qualitative and quantitative performance compared to prior and contemporary methods. We show significant improvements from prior work [DLJ*20, SYMB21], in terms of robustness, ability to model complex clothing styles, and retaining fine pose-dependent details. The use of generalized implicit functions with our

canonical mapping and displacement field networks helps the network to more effectively factor out articulation from non-rigid components. We also extend our model to multiple shape setting and show comparable performance to concurrent work LEAP [MZBT21].

In the following sections, we first discuss the limitations of NeuralGIF due to its underlying formulation and review contemporary works. Finally, we conclude with the impact of NeuralGIF and offer some insights for future work.

5.7.1 Limitation and Discussions

Despite its strengths, Neural-GIF has notable limitations. The current implementation is clothing-specific, primarily designed and trained for a single subject wearing a fixed cloth. Specifically given multiple scans of a person in clothing, NeuralGIF learns to animate the person in new poses. This specificity restricts the method's applicability in scenarios requiring diverse clothing styles or simultaneous modeling of multiple garments, such as varying clothing types across individuals or multiple subjects with multiple clothing. More recent works, such as gDNA [CJS*22], have addressed these constraints by extending neural implicit representations into multi-cloth and multi-shape frameworks, enabling more versatile applications and broader generalization across different clothing types.

Neural-GIF's backward skinning formulation, while effective, inherently limits its generalization to extreme, previously unseen poses. The network requires careful pre-training with pose-specific data to perform well in highly dynamic or rare articulations, making pose generalization challenging without extensive and diverse training datasets. NeuralGIF formulation exclusively captures geometric deformation, neglecting texture and appearance details. The addition of texture modeling, potentially utilizing strategies similar to those employed by SCANimate [SYMB21], could increase the applicability of NeuralGIF.

Neural Implicits for human in clothing models

Although significant progress has been made in improving the quality of results [TSTPM21, SYMB21], pose generalization [CZB*21, MSB*22], multi-subject modeling [MSB*22, MZBT21, CJS*22], and dynamic modeling [BBR*22], there remains a substantial gap between these models and their practical application in real-world tasks. Several works have adopted similar formulations in NeRF-based setups [WCS*22, ZYZ*22, PDW*21, IRG*23b]. Implicit methods typically involve computationally intensive inference steps and are incompatible with traditional graphics pipelines. This computational cost restricts their use in real-time applications and integration with conventional graphics workflows, which often favor mesh-based or explicit representations.

Recent advancements have partially mitigated these limitations by introducing more efficient rendering techniques tailored to implicit surfaces [VSJ22], such as Signed Distance Function (SDF)-based renderers, which improve compatibility and reduce computational overhead. However, a newer representation based on 3D Gaussian Splatting (3DGS)[KKLD23] has recently begun to supersede implicit representations. Building on prior research into skinning and reposing neural implicits, several works have transferred this knowledge to 3DGS-based representations[RASS24, QWM*24, UFF06, JBS*23].

Despite these advances, implicit methods inherently struggle to model open surfaces and multilayered structures. They usually represent clothing and the human body as a continuous, closed surface, which limits their ability to accurately capture clothing with complex layering or open geometries, such as coats or skirts. Several works have employed unsigned distance fields (UDFs) to represent open surfaces like clothing [CMPM20, CAPMMN24, CPA*21a]. More recently, some approaches have proposed differentiable methods to reconstruct meshes from UDFs [GSF22, DLLG*23, LGF23]. These limitations motivate us to explore a more explicit representation for modeling clothing or humans in clothing, which we visit in the next chapter.

5.7.2 Future Work and Impact

Neural-GIF has notably influenced subsequent research, motivating important follow-up methods such as [WCS*22, PDW*21], which extend our backward skinning formulation to achieve improved realism and dynamic human modeling capabilities. HumanNeRF, for instance, leverages implicit volume rendering methods inspired by our approach, enabling more realistic modeling and animation of human subjects, including dynamic appearance changes such as viewpoint-dependent textures and reflections. The ClothSeq dataset introduced in our work remains highly valuable for modeling complex clothing dynamics and has spurred further advancements [QXL*22], with subsequent datasets, such as 4D-Dress, introducing even more comprehensive scans that capture detailed temporal variations and intricate clothing interactions.

Neural Generalized Implicit Functions open several interesting research directions. Neuralgif is a clothing-specific model; it would be useful to extend this approach such that it can animate multiple clothing items using the same model. This has been addressed in some works [CJS*22], but they employ a forward skinning approach and thus yield more rigid results. It would also be valuable to learn temporal correspondences implicitly during the learning process. Finally, since Neural-GIF produces signed distance fields, it can also be used for fast collision and contact computation in human-object and scene interactions, as done by COAP [MSB*22]. However, this approach only works for unclothed humans.

Furthermore, the recent emergence and rapid advancement of point-based and 3D Gaussian Splatting (3DGS)-based representations mark a significant shift away from traditional neural implicit methods. These approaches offer significantly improved computational efficiency, scalability, and better compatibility with established graphics pipelines, making them attractive for both research and practical applications in real-time environments such as gaming and augmented reality. The integration of neural implicit insights with these newer representation frameworks promises exciting future developments. We anticipate ongoing research will focus on combining the expressive power and flexibility of neural implicit models with the efficiency of point-based and 3DGS techniques, thus creating hybrid representations that deliver even faster, more versatile, and visually compelling solutions for realistic human and clothing animation.

6

TailorDiff: Unified and Expressive Generative Model for 3D Clothing

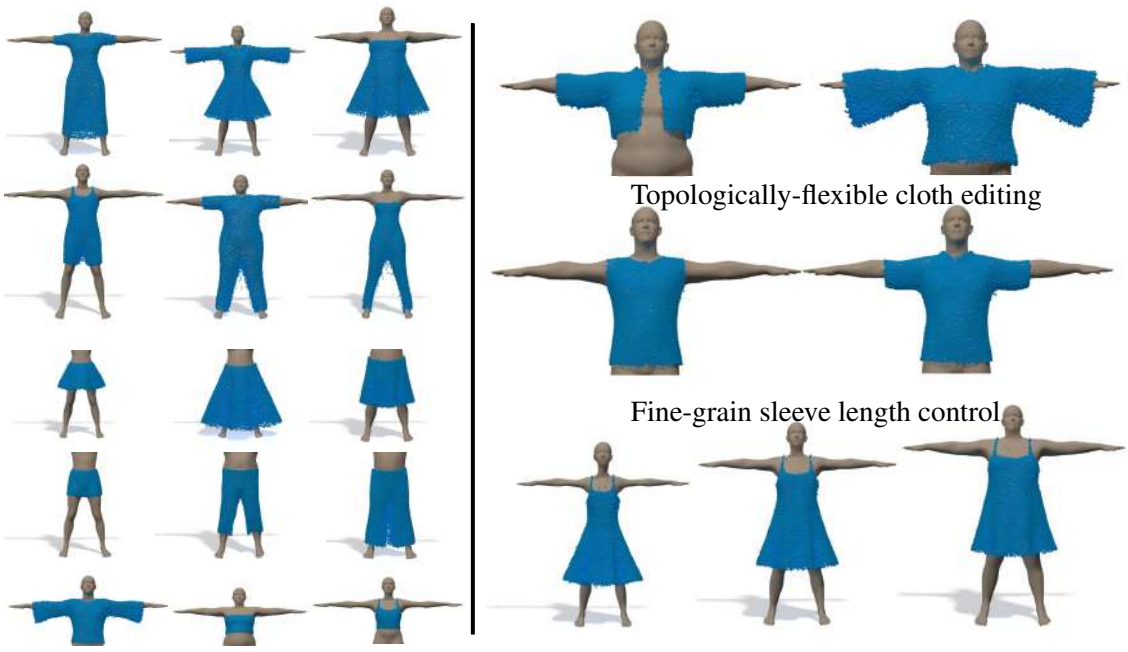


Fig. 6.1 **TailorDiff- A unified model of clothing:** We present TailorDiff, a hierarchical latent point diffusion model for unified and expressive generative modeling of 3D clothing. TailorDiff is topologically flexible and generates diverse clothing for diverse body shapes with a single model (*left*). The hierarchical and body-aware latent space of TailorDiff provides coarse to fine-grained control over generation (*right*).

Realistic and editable modeling of human avatars is essential for a wide range of applications, including telepresence, entertainment, and virtual try-on systems. Despite recent advances, most existing generative methods represent clothing and the human body as a single monolithic surface. This design choice significantly limits their ability to realistically capture layered clothing and its interactions with the underlying body.

We introduce TailorDiff, a Hierarchical Latent Point Diffusion Model designed for unified, expressive, and high-fidelity generative modeling of 3D clothing. To the best of our knowledge, TailorDiff is the first model that can represent clothing with arbitrary topology

within a single, coherent framework. It enables the generation of a wide variety of clothing geometries conditioned on several attributes such as clothing style, underlying body shape, and fine-grained semantic controls like sleeve length, hem length, and more. A key idea in our approach is a novel body-aware encoding that ensures the generated garments are well-aligned with the human body and conform to realistic layering constraints between the body and clothing. As a result, TailorDiff is not only a powerful generative model but also highly applicable to inverse problems—for instance, fitting a human-in-layered-clothing representation to partial or occluded observations such as 3D scans or monocular images. We demonstrate the capabilities of TailorDiff in Figure 6.1.

This chapter is based on our ongoing, unpublished work titled TailorDiff: A Unified and Expressive Generative Model for 3D Clothing. The overarching goal of this work is to develop a unified clothing model that is inherently compatible with existing human body representations, thereby facilitating a broad range of applications, particularly in solving inverse problems from partial observations such as images, videos, and scans. The emergence of such models will drive the field toward a truly unified and layered representation of humans and their clothing, one that more accurately captures real-world complexity. Such representations can serve as a robust representation for tasks involving human–cloth interaction, physically plausible simulation, and other applications.

6.1 Introduction

Realistic and editable modeling of humans wearing diverse clothing is crucial for immersive applications such as telepresence, entertainment, virtual try-ons, augmented and virtual reality experiences, and digital human understanding. SMPL [LMR*15] has been instrumental in understanding and reconstructing human bodies from partial observations such as images, videos, or scans. However, SMPL primarily models the naked human form or, at best, clothing as simplistic offsets from the underlying body mesh. It inherently lacks the capacity and ease to represent complex clothing structures, multi-layered outfits, and garments with varying topologies and intricate geometries. This limitation substantially restricts its expressiveness and applicability in real-world scenarios, where detailed clothing representation is essential. As we have discussed throughout the previous chapter, significant progress has been made to represent clothed humans, broadly categorized into mesh-based and neural implicit-based methods. Mesh-based methods, while beneficial for inverse problems [MYR*20, PLPM20] (e.g., fitting clothing models to partial observations), typically rely on predefined templates [TBTPM20] or are limited to body conforming geometries [MYR*20]. These templates limit the flexibility in representing clothing that diverges significantly from the SMPL topology, necessitating the manual creation of multiple garment templates for each distinct clothing type. Consequently, mesh-based methods struggle with generalization, limiting their applicability to a diverse range of clothing styles and categories. On the other hand, neural implicit methods have demonstrated impressive capabilities in modeling complex clothing geometries that deviate significantly from the underlying body topology, without the constraints of predefined templates. Some works have also shown fitting neural implicit-based representation to scans [CJS*22, CPA*21a], images [DLLG*23, LGF23] *etc.*.

Another key aspect of the human in clothing model is the disentanglement of the body and different clothing layers.

Nevertheless, most prior approaches fall short in handling multi-layered clothing structures. They frequently merge clothing and body representation into a single implicit function or single mesh layer, thereby losing the critical distinction and interaction between individual clothing layers and the underlying body. Such combined representations complicate editing, reusability, and physical realism, essential aspects for applications like virtual try-on and clothing animation. Modeling clothing and the human body together fails to capture the detailed thin structures and layered nature of real-world clothing, and the clothing cannot be reused for applications such as virtual try-on. Some works like [BTTPM19, JZH*20] predict human and clothing as separate meshes, given single or multi-view images. However, they are still limited by predefined clothing categories. Neural Implicit-based works [CPA*21a, DLLG*23] propose a fitting pipeline to recover layered human+ clothing from images given their segmentation map. However, these methods come with an inherent problem of accurately representing open surfaces using a distance field. [CPA*21a] results in coarse geomtery, while [DLLG*23] sometimes results in unstable results. Moreover, these models are still defined for limited clothing and need different models for upper, lower, and other body parts. Another line of work, such as DELTA [FLB*23] uses hybrid approaches to model disentangled human and clothing using a mesh-based representation and a NeRF-based representation for clothing and hair. We provide a detailed overview of prior clothing models in Table 6.1. As observed from Table 6.1, despite impressive visuals, none of these families of models offers the universal solution required for SMPL-level impact:

- **Category-agnostic diversity:** Ability to sample T-shirts, hoodies, skirts, suits, etc., from a single latent space;
- **True disentanglement:** Separate, editable latent codes for body shape/pose and for each garment layer, enabling mixing-and-matching and layering;
- **Inverse-problem readiness:** A differentiable, fast, parametrized model that can be optimized against sparse observations (single images, videos, scans, depth scans, or point clouds), just as SMPL is today.

Because existing clothing models violate at least one of these criteria, they have not become the de facto backbone for reconstruction, completion, or animation in the way SMPL has for naked bodies. In this work, we introduce a unified model for clothing that can produce diverse and realistic clothing, is highly controllable, and is capable of fitting to partial observations such as multi-view images and scans. It enables disentangled and layered representation of *humans in clothing*, laying the foundation not just for photorealistic, but also physically realistic and editable avatars.

In this work, we address these limitations by proposing a unified and highly controllable clothing model explicitly designed to integrate seamlessly with the SMPL body model. Our model aims to produce diverse and realistic clothing, independently modeling each clothing layer and hence enabling the complex geometrical and physical interactions among garments

Method	Body shape	Flexible topology	Category-agnostic	Fine-grain control	Correspondence model
DIG [RGRF22]	✗	✗	✗	✗	✗
DrapeNet [DLLG*23]	✗	✗	✗	✗	✗
SMPLicit [CPA*21a]	✓	✓	✗	✓	✗
ISP [LGF23]	✗	✓	✗	✗	✗
Neural-ABC [HYZ24]	✓	✓	✗	✗	✗
WordRobe [SMR*24]	✗	✓	✓	✓	✗
TailorDiff	✓	✓	✓	✓	✓

Table 6.1 **TailorDiff v/s prior works for unified model of clothing:** Comparison of our method with other works. Ours is the first unified model for generating clothing of all topologies and categories using a single model.

and between clothing and the human body. By disentangling clothing from the body and enabling layered representation, our model inherently accommodates partial observations, occlusions, and observational noise, critical for real-world reconstruction scenarios. Unlike prior models, our unified approach does not rely on restrictive clothing templates or predefined garment categories. Instead, it can interpolate and generate diverse clothing geometries and topologies, facilitating rich variations in style, shape, and layering. We show results of interpolating between different topologies in Figure 6.1. This generalization makes our model uniquely suited for inverse problem-solving tasks such as fitting clothing to incomplete multi-view images, monocular videos, scans, or noisy observations, thereby enabling more accurate and realistic reconstructions of clothed humans. Our model unlocks new possibilities in digital human modeling, significantly broadening the scope and applicability of virtual avatar creation, human digitization, virtual try-ons, and animation.

Developing such a comprehensive clothing model presents several challenges. Firstly, training a highly expressive large-scale model is inherently difficult due to the diversity and complexity of clothing styles and topologies. Secondly, capturing this diversity requires a large-scale dataset, and there is no single large-scale dataset of diverse clothing available. Moreover, existing smaller datasets are not always aligned with a common body model or pose, making it difficult to learn a unified representation. Finally, incorporating the body shape information into the model to respect physical constraints on the layered clothing is non-trivial, especially when we need a model that works for all genders, unlike prior works [CPA*21a, PLPM20, LGF23] that are gender-specific and accommodate a wide variety of body shapes.

To address these challenges, we propose TailorDiff, a large-scale and unified generative model for 3D clothing, based on point clouds. Our motivation for using point cloud representations stems from the topological flexibility of point clouds [MYTB21, PMR*23] TailorDiff is modeled as a hierarchical latent diffusion model, which enables the model to learn detailed and structured latent representations crucial for modeling the complexity of clothing shapes. TailorDiff can be used to generate diverse clothing for a wide array of diverse body shapes (see Figure 6.1). Our model can sample any clothing category of varying topology and interpolate between different clothing items. We address the lack of a large-scale, diverse

dataset by combining multiple high-quality, publicly available datasets of diverse clothing items [KKK*24, BME20, ZHW23] into a unified, large-scale dataset, aligning all items to SMPL [LMR*15]. We use the dataset to train a highly expressive large-scale model by leveraging a hierarchical VAE structure combined with Denoising Diffusion Models (DDMs) operating in the latent space. The hierarchical latent point diffusion model offers fine-grained control over clothing generation and generalizes effectively across various garments, body shapes, and genders. Furthermore, to incorporate underlying body information, we introduce a structured body latent representation that explicitly encodes the human body shape within the model. This ensures that the generated clothing adapts to different body types, significantly improving alignment and reducing intersections between clothing and the body.

We show application of TailorDiff in clothing generation and editing, achieving impressive results that surpass previous works. Our model not only produces a greater variety of clothing styles but also provides flexible control from coarse to fine-grained detail, opening up new creative possibilities for virtual avatar creation. The latent space of TailorDiff is both robust and reliable, consistently generating realistic clothing. Additionally, our body-aware encoding enhances the alignment between body and clothing, significantly reducing intersections. We also demonstrate the versatility of TailorDiff for inverse problems, by effectively fitting to scans. In summary, our contributions are:

- We present TailorDiff, the first large-scale generative model for 3D clothing, which encapsulates diverse clothing styles in a unified framework.
- We incorporate explicit body encoding into our model, which enables better body-cloth alignment and respects correct layering. Moreover, it works for all genders, overcoming the limitations of gender-specific models.
- We design a hierarchical and structured cloth latent space using diffusion models, resulting in fine-grained control over cloth modeling.

We begin by introducing TailorDiff in Section 6.2, where we describe the hierarchical latent space and the diffusion-based prior learned over it. To train this large-scale generative model, we leverage several publicly available datasets of synthetic clothing items, with details on dataset scale and preprocessing discussed in Section 6.3. We then present extensive experiments and results in Section 6.4, demonstrating the capabilities of our learned prior across various tasks, including editing, generation, and fitting. Finally, in Section 6.5, we conclude with a discussion on the limitations of point cloud-based representations and share our insights on building a unified model for humans and clothing.

6.2 A Unified and Expressive Generative Model for 3D Clothing

In this section, we introduce TailorDiff, a hierarchical latent point diffusion model for unified and expressive generative modeling of 3D clothing. TailorDiff is structured as a hierarchical Variational Autoencoder (VAE) with Denoising Diffusion Models (DDMs) operating in

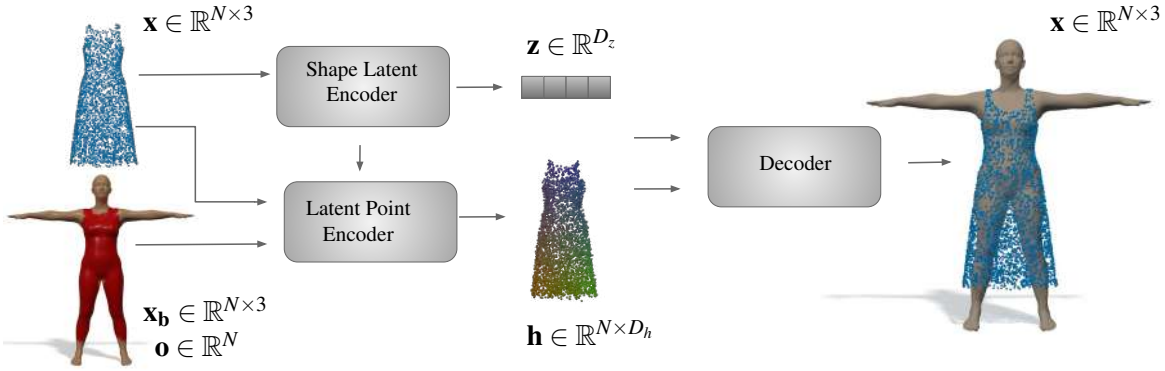


Fig. 6.2 **TailorDiff Hierarchical VAE:** TailorDiff is structured as a hierarchical VAE with DDM operating in the latent space. Given a clothing pointcloud $\mathbf{c}$ with corresponding body shape ($\mathbf{x_b}$) clothing occupancy on body($\mathbf{o}$), a VAE model is trained which learns a global latent shape ($\mathbf{z} \in \mathbb{R}^{D_z}$) and a local point latent ($\mathbf{h} \in \mathbb{R}^{N\times D_h}$).

the latent space [RBL*21, ZVW*22]. The complexity of clothing shapes presents unique challenges due to their diversity, intricate structures, and fine-grained details. Previous methods [LH21b] that rely solely on global shape encoding have struggled to capture fine details in common objects, making them less suitable for modeling clothing. To address this, we propose to model the latent space of TailorDiff using a global shape latent representation with a body-aware, point-structured latent space. The hierarchical latent space enhances the model's generation capability by learning a detailed and structured latent representation, as opposed to other works, such as [LH21b, ZDW21]. Additionally, the body-aware VAE implicitly models the relationship between the body and clothing, thereby contributing to better alignment between the body and clothing. We provide an overview figure of our method in Figure 6.2 and Figure 6.3 . We train TailorDiff, by first learning the hierarchical body-aware VAE encoder-decoder (see Section 6.2.1), and then train Denoising Diffusion Models on the resulting latent space in (see Section 6.2.2). Once we have unified clothing prior learned, we can use it for conditional generation. We discuss conditional generation using TailorDiff in Section 6.2.2.1.

6.2.1 Hierarchical Latent Space

We first train a hierarchical VAE using large-scale point cloud-based clothing data and a fitted SMPL body. Given a clothing shape as point clouds $\mathbf{x} \in \mathbb{R}^{N\times 3}$, we design a body-aware VAE to learn a hierarchical latent space. This hierarchical latent space utilizes two levels:

1. **Global Shape Latent $\mathbf{z}_0 \in \mathbb{R}^{D_z}$:** Captures the coarse and overall style of clothing.
2. **Latent Point Cloud $\mathbf{h}_0 \in \mathbb{R}^{N\times D_h}$:** Encodes fine-grained details—including the relationship to the body and small clothing features such as sleeve length—and is well-suited for diffusion-based point cloud generation.

To learn two-level latent codes in a disentangled manner, we define the encoder as a combination of a *Shape Latent Encoder* g_ψ and a *Latent Point Encoder* q_ϕ. The global

Shape Latent Encoder encodes the point cloud into the global shape latent vector $\mathbf{z}_0$. This global shape latent $\mathbf{z}_0$ along with a clothing class embedding $\mathbf{c}$ is then concatenated with each point in the point cloud to learn the per-point features using a *Latent Point Encoder*. Incorporating body information into the model is essential for aligning clothing with the body. Prior works [BTTPM19, MYR*20, DLLG*23] have modeled this by predicting clothing as an offset from the body; however, this approach limits the model's ability to represent clothing items such as dresses and skirts. Moreover, predicting clothing as an offset may not effectively capture the complex interactions between cloth and body, which are essential for realistic clothing generation. We propose to incorporate body-cloth interaction by learning body-aware features in the per-point latent space. We will introduce our novel body-aware feature encoder next and explain the modified latent space.

Body-Aware Encoding: SMPL shape parameters (β) are commonly used to encode body information into models, but they make the model gender-specific. Instead, we represent body shape using an explicit point cloud, which overcomes the limitation of gender specificity. Specifically, we represent body information using $\mathbf{x}_b$, where $\mathbf{x}_b \in \text{SMPL}(\beta, \theta_0)$ and $\mathbf{o} \in \mathbb{R}^{N\times 1}$, where $o_i = 1$ if the body point is occupied by clothing and $o_i = 0$ otherwise. This concatenated input is then passed to q_ϕ, which learns per-point features influenced by both clothing geometry and body context.

To incorporate the proposed body input into the model, we introduce a novel body-aware encoding based on PVConv [LTLH19]. PVConv effectively combines point-based features using MLPs with the strong spatial inductive bias of convolutions through Voxel-CNN. We hypothesize that, though body information is necessary for alignment and interaction, per-point local features are not required; instead, a coarser shape representation of the body can efficiently encode this information. Hence, we modify the original PVConv layer by introducing an additional Voxel-CNN layer to encode body information. We refer to this modified layer as Body-PVConv. So Body-PVConv has three components, one MLP f_p, and two Voxel-CNN layers f_vc and f_vb. The layer is defined as:

$$f_\text{PV}(\mathbf{x}_i) = f_\text{p}(\mathbf{x}_i) + f_\text{vc}\left(V(\mathbf{x}_i)\right) + f_\text{vb}\left(V(\mathbf{b}_i)\right), \tag{6.1}$$

where f_p is a per-point MLP applied to $\mathbf{x}_i$, f_vb denotes the voxel convolution operation on body features $\mathbf{b}_i$ and f_vc denotes the same on cloth features, $V(\mathbf{x}_i)$ represents the voxelization of the neighborhood of point $\mathbf{x}_i$. Trilinear interpolation is used to map voxel features to per-point features. This constitutes our single Body-PVConv layer. We utilize a PVConv [LTLH19]-based architecture to implement g_ψ and Body-PVConv to implement q_ϕ and p_ξ, as it takes body information as input. Finally, the encoders and decoders of our body-aware hierarchical VAE are:

$$\begin{aligned}
\hat{\mathbf{z}}_0 &\sim g_\psi(\mathbf{z}_0 \mid \mathbf{x}),\\
\hat{\mathbf{h}}_0 &\sim q_\phi(\mathbf{h}_0 \mid x, \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c}),\\
\hat{\mathbf{x_c}} &\sim p_\xi(x \mid \mathbf{h}_0, \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c})
\end{aligned} \tag{6.2}$$

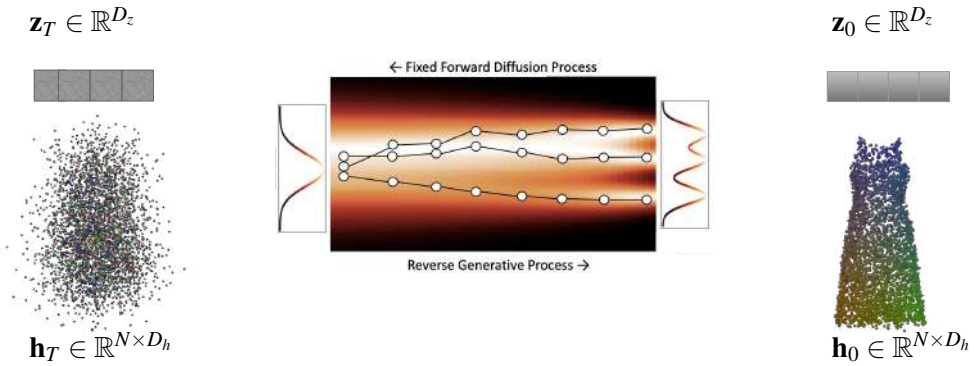


Fig. 6.3 **TailorDiff Latent Diffusion Model**: A DDM is trained on the learned clothing latent space. TailorDiff results in diverse and coarse-to-fine-grained control over clothing generation using the learned latent and input conditions.

We train the VAE following a similar training regime, as mentioned in [ZVW*22]. After training the VAE, we have a regularized hierarchical latent space. Next, we learn a highly expressive latent diffusion model on this latent space.

6.2.2 Latent Point Diffusion Model

Following the approach in [ZVW*22], we train DDMs on the learned regularized latent space. This is motivated by the observation that VAE priors exhibit prior hole problems [TW17, VKK21] and hence are not highly accurate in matching the training distribution. Specifically, TailorDiff learns two latent DDMs ε_ψ and ε_ϕ on the latent spaces $\mathbf{z}_0$ and $\mathbf{h}_0$ using the objectives given in Equation (6.3) and Equation (6.4), where $\varepsilon \sim \mathcal{N}(0,I)$ and $t \sim \mathcal{U}\{1,T\}$.

$$\mathcal{L}^{\mathbf{z}}_{\text{DDM}}(\psi) = \mathbb{E}_{p(x),q_\phi(\mathbf{z}_0|x)} \left\| \varepsilon - \varepsilon_\psi(\mathbf{z}_t, \mathbf{c}, t) \right\|^2, \tag{6.3}$$

$$\mathcal{L}^{\mathbf{h}}_{\text{DDM}}(\phi) = \mathbb{E}_{\substack{p(x),q_\phi(\mathbf{z}_0|x),\\ q_\phi(\mathbf{h}_0|x,\mathbf{z}_0,\mathbf{x_b},\mathbf{o},\mathbf{c})}} \left\| \varepsilon - \varepsilon_\psi(\mathbf{h}_t, \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c}, t) \right\|^2, \tag{6.4}$$

The diffused latent codes in Equation (6.3) and Equation (6.4) are given by $\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \varepsilon$ and $\mathbf{h}_t = \alpha_t \mathbf{h}_0 + \sigma_t \varepsilon$. We learn conditional models that can be controlled by body shape, clothing occupancy on the body, and clothing class. This makes TailorDiff highly controllable and expressive. While training the DDMs, we keep the VAE encoder and decoder fixed.

6.2.2.1 Conditional generation from learned TailorDiff

In the generation phase, the latent denoising diffusion models define a hierarchical generative model:

$$x \sim p_\xi(x \mid \mathbf{h}_0, \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c})\, p_\phi(\mathbf{h}_0 \mid \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c})\, p_\psi(\mathbf{z}_0 \mid \mathbf{c}). \tag{6.5}$$

We sample the latent DDMs hierarchically and decode the latents back to the original point cloud space using p_ξ. We can use TailorDiff for clothing generation, editing by manip-

ulating $\mathbf{h}_0, \mathbf{z}_0, \mathbf{x_b}, \mathbf{o}, \mathbf{c}$.

Latent representation for o: To control generation using the occupancy of clothing on the body, we train a simple MLP-based autoencoder, which learns to map a latent code $\mathbf{z_c}$ to $\mathbf{o}$. This facilitates the user to easily generate new styles by interpolation in the latent space of $\mathbf{z_c}$.

Implementation Details: We implement a VAE latent point encoder and decoder using 3 layers of Body-PVConv and a shape latent encoder using 2 layers of PVConv. Shape latent DDM is implemented using ResNet, and latent point DDM is implemented using PVConv layers.

6.3 Dataset

6.3.1 Synthetic Clothing Datasets

Prior work modeling *clothing* or *humans in clothing* has mainly relied on individual synthetic datasets for training, which often exhibit limited diversity in clothing styles, body shapes, and genders. For example, WordRobe [SMR*24] used clothing style variation data [KL21] with no body shape variation. Similarly, DrapeNet [DLLG*23] utilizes a small subset of Cloth3D, which has limited clothing style variation and only female bodies. Consequently, prior work relies on small datasets with limited variation, impacting the model's expressivity in capturing the diversity and complexity of clothing and its ability to generalize draping across different body shapes. We enhance the training data by combining multiple high-quality, large-scale, publicly available synthetic datasets. Specifically, we use Cloth3D [BME20], GarmentCodeData [KKK*24], Cloth4D [ZHW23], and ClothesNet [ZZL*23]. This results in a dataset comprising 68*k* clothing items covering a wide range of clothing. In contrast to prior work, which relies on gender-specific datasets, predominantly featuring neutral [KKK*24] or female [ZHW23] body types, our combined dataset includes male, female, and neutral body types. Moreover, since our model does not utilize SMPL gender-dependent shape parameters but instead relies on the underlying body geometry to generate clothing, it is also gender-agnostic. We use 3 clothing classes, namely upper (T-shirt, jacket, *etc.*), lower (skirt, pants, *etc.*), and combined (dress, jumpsuits, *etc.*), and define a class embedding for each clothing class.

As mentioned earlier, prior works have predominantly relied on a single data source for training their models. This limits diversity across various aspects such as clothing style, classes, size *etc.*. For instance, WordRobe [SMR*24] is trained exclusively on a clothing style variation dataset (IGL [KL21]) and cannot model shape-dependent deformations. Similarly, methods like DrapeNet, DIG, and ISP [BME20] incorporate body shape variations but are trained on datasets with a limited range of clothing styles.
In contrast, we propose learning from a large-scale dataset that encompasses variations in body shape and clothing style by consolidating multiple datasets with a diverse array of body shapes and clothing styles. To this end, we utilize Cloth3D [BME20], Cloth4D [ZHW23],

Dataset	# of Clothing	SMPL	Diverse Clothing Class	# Final curated samples
Cloth3D [BME20]	$\sim 11k$	✓	✓	$\sim 11k$
Cloth4D [ZHW23]	400	✗	✗(dresses)	~ 350
GarmentCodeData [KKK*24]	$\sim 125k$	✗	✓	$\sim 56k$
ClothesNet [ZZL*23]	2277	No body alignment	✓	1397

Table 6.2 **Dataset Overview:** Details of publicly available datasets used for training our model. We combine all these datasets by aligning them to a single body model (SMPL in our case) and posing all the clothing scans in a fixed canonical pose. Finally, the samples mentioned in the last column are high-quality, well-aligned scans that are used to train and evaluate our model.

GarmentCodeData [KKK*24], and ClothesNet [ZZL*23]. We provide more details about these datasets in table Table 6.2. Each of these datasets offers distinct advantages, such as diversity in clothing styles and body shapes. However, they cannot be readily used to train a model, as they are designed/captured in different poses. For example:

- **Cloth3D** is captured in a TA-pose (*i.e.*, hands in a T-pose and legs in an A-pose) and provides SMPL parameters.
- **Cloth4D** contains 4D scans with clothing in an A-pose. It provides a body-under-clothing mesh derived from SMPLX. However, this mesh is combined with sub-meshes, such as hair and shoes, which alter the mesh's topology. Moreover, SMPLX parameters are not provided. Therefore, making this dataset consistent with others in terms of pose and underlying body information requires additional processing.
- **GarmentCodeData** uses a non-public body model with parameters and topology that are incompatible with the SMPL body model.
- **ClothesNet** does not include any underlying body information at all.

These discrepancies make it challenging to use all datasets within a unified framework. To address this, we align all datasets to the SMPL body model and pose them to a common canonical pose. In the following section, we discuss the pre-processing steps of all datasets.

6.3.2 Dataset Processing and Alignment

In this section, we provide a detailed overview of the datasets used in our work to curate a larger dataset with diverse clothing, gender and body shape. We also describe the processing pipeline employed to clean, align, repose and integrate these datasets into a unified format suitable for downstream tasks.

Cloth3D [BME20]: The Cloth3D dataset consists of a large collection of artist-designed clothing items. Since our focus is on static and fixed canonical poses, we only utilize the static scans from the dataset. Cloth3D includes SMPL pose and shape parameters. Cloth3D

scans are in TA pose, and we use this pose as our canonical pose. Additionally, these SMPL parameters allow us to extract the body geometry underneath the clothing. We make use of all the static scans provided in this dataset, and finally, we have $\sim 11k$ clothing items from 5 clothing classes in this data subset.

Cloth4D [ZHW23]: Cloth4D dataset consists of high-quality, texture-rich synthetic 4D scans containing diverse clothing styles (mainly dresses). This dataset contains details about hair, facial features, accessories such as shoes, and some intricate clothing designs. Although the dataset comes with detailed body under clothing mesh, it doesn't provide well-aligned SMPL parameters, and also the *body under clothing* doesn't share the same topology as SMPL or SMPL-X. In other words, Cloth4D does not provide SMPL or SMPL-X parameters. We devise an SMPL registration pipeline to fit the body underneath and define our optimization using the following:

$$\mathcal{L}(\beta,\theta) = \mathrm{w}_{\text{data}}\mathcal{D}(M(\beta,\theta),\mathcal{S}) + \mathrm{w}_{\beta}||\beta|| + \mathrm{w}_{\theta}||\theta - \theta_{\text{init}}|| \tag{6.6}$$

where β,θ are SMPL shape and pose parameters. $\mathcal{D}$ represents the bidirectional Chamfer distance between the SMPL mesh M and the scan surface $\mathcal{S}$. The other two terms are prior terms for body shape and a penalty term for body pose. The body shape prior ensures that the body shape does not deviate significantly from the average body shape, while the pose penalty term ensures that the body pose does not deviate excessively from the initial pose θ_{init}. Since we know that the first frame of Cloth4D resembles an A-pose, we initialize the pose as an A-pose and penalize large deviations from this initial value. Moreover, as the scan surface already represents the human body without clothing, this becomes a relatively simple optimization problem.

Cloth4D consists of 400 clothing items. We apply our method to the available scans and filter out those for which the clothing is too complicated. We further change of pose of clothing to TA-pose. In the end, we are left with ~ 350 scans.

GarmentCodeData [KSH23, KKK*24]: The GarmentCode Dataset is a large-scale clothing dataset comprising clothing designs created using sewing patterns. The dataset contains diverse clothing designs with shape variation. However, the body model provided in this dataset is not compatible with SMPL. Hence, we register the dataset with the SMPL body model. Since accurate *body under clothing* is provided, we use the optimization proposed for Cloth4D dataset in Equation (6.6) to get SMPL parameters from the given body.

ClothesNet [ZZL*23]: ClothesNet is a diverse clothing style dataset of limited clothing items and some accessories. We only deal with clothing items, as this is our interest. The underlying body is not provided for this data. Thus, we need to register SMPL to obtain the body. This is a much harder problem, as clothing meshes can only reflect partial information about the body. We use TailorNet [PLPM20] to initialize the rough body fitting for each

clothing. Specifically, we do the following initialization:

$$\begin{aligned}\mathcal{L}(\beta,\theta,\gamma) = \mathrm{w}_{\text{data}}\mathcal{D}(f_{\text{TN}}(\beta,\theta,\gamma),\mathcal{S}) + \mathrm{w}_{\beta}||\beta|| \\ +\mathrm{w}_{\gamma}||\gamma|| + \mathrm{w}_{\theta}||\theta-\theta_{\text{init}}||\end{aligned} \tag{6.7}$$

where, similar to Equation (6.6), β,θ are SMPL shape and pose parameters, γ is the clothing style parameter of TailorNet and f_{TN} represents the TailorNet model, which outputs a mesh given β,θ,γ as input. We use class-specific models for this optimization. For dress, we optimize the model for both skirt and t-shirt together, with shared β,θ parameters. After this initialization, we use the following optimization to get the body under clothing:

$$\begin{aligned}\mathcal{L}(\beta,\theta) = \mathrm{w}_{\text{data}}\mathcal{D}_{\mathcal{S}\to\mathrm{M}}(M(\beta,\theta),\mathcal{S}) + f_{\text{in}}(M,\mathcal{S}) \\ +\mathrm{w}_{\beta}||\beta-\beta_{\text{init}}|| + \mathrm{w}_{\theta}||\theta-\theta_{\text{init}}||\end{aligned} \tag{6.8}$$

where, $\mathcal{D}_{\mathcal{S}\to\mathrm{M}}$ is unidirectional chamfer distance from clothing surface to SMPL body surface. $f_{\text{in}}(M,\mathcal{S})$ is the term which penalises if the points of clothing surface lie outside the body mesh and is given by :

$$f_{\text{in}}(M,\mathcal{S}) = \begin{cases} 0 & \text{if}, g(p,M) > 0 \\ |d(p,M)| & \text{otherwise.}\end{cases} \tag{6.9}$$

where, p are the points on the clothing surface $\mathcal{S}$ and g is occupancy value obtained from COAP [MSB*22], which is an implicit representation for SMPL body model. Since the poses we use are very simple, COAP predictions are reliable and robust.

6.4 Experiments and Results

In this section, we provide extensive results for our proposed model and its applications. In Section 6.4.1, we show the ability of TailorDiff for clothing generation and editing, and compare with relevant prior works. We further show the application of our model for *Fitting to scans*. We conclude with ablation studies on our proposed model, experimentally validating our design choices.

6.4.1 Generation using TailorDiff

Using TailorDiff, we can generate clothing samples by denoising $\mathbf{h}$ and $\mathbf{z}$. Since the model is trained as a conditional model, we can control the generation by modifying the input control signals $\mathbf{c}$ and the body encoding $\mathbf{x_b}, \mathbf{z_c}$, *i.e. body shape* and *clothing occupancy on body*. As seen in Figure 6.1, TailorDiff can generate diverse and topologically flexible clothing shapes from a single unified model.

Quantitative Evaluation We compare the results of clothing generation with prior works such as NeuralABC [HYZ24], ISP [LGF23], DIG [RGRF22], DrapeNet [DLLG*23], and

	Upper Clothing							Lower Clothing						
	JSD(↓)	COV(%↑)		MMD(↓)		1-NNA(%↓)		JSD(↓)	COV(%↑)		MMD(↓)		1-NNA(%↓)	
		CD	EMD	CD	EMD	CD	EMD		CD	EMD	CD	EMD	CD	EMD
SMPLicit [CPA*21a]	10.182	40.0	34.5	0.588	0.530	92.5	89.0	11.569	36.3	34.9	0.275	**0.198**	93.0	84.5
DIG [RGRF22]	5.319	43.9	30.1	0.325	0.341	86.0	87.0	7.354	42.2	45.3	0.249	0.228	73.5	64.5
DrapeNet [DLLG*23]	10.126	40.3	41.8	0.316	0.293	84.0	67.0	8.218	50.5	45.6	0.305	0.295	86.5	78.0.5
ISP [LGF23]	7.625	39.6	40.2	0.339	0.334	90.5	82.0	9.275	36.2	327.3	0.343	0.316	96.0	91.5
TailorDiff (Ours)	**2.106**	**50.1**	**50.8**	**0.192**	**0.166**	**67.6**	**63.5**	**1.968**	**54.3**	**56.2**	**0.218**	0.232	**58.7**	**56.5**
	Skirts							Dresses/Jumpsuits						
	JSD(↓)	COV(%↑)		MMD(↓)		1-NNA(%↓)		JSD(↓)	COV(%↑)		MMD(↓)		1-NNA(%↓)	
		CD	EMD	CD	EMD	CD	EMD		CD	EMD	CD	EMD	CD	EMD
SMPLicit [CPA*21a]	25.532	30.4	23.5	**0.426**	**0.300**	96.5	86.5	-	-	-	-	-	-	-
DIG [RGRF22]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DrapeNet [DLLG*23]	-	-	-	-	-	-	-	-	-	-	-	-	-	-
NeuralABC [HYZ24]	-	-	-	-	-	-	-	10.198	49.5	54.6	0.829	0.682	**62.5**	**59.5**
TailorDiff (Ours)	**6.149**	**48.4**	**54.5**	0.574	0.345	**68.3**	**58.8**	**3.172**	**52.8**	**55.6**	**0.578**	**0.546**	63.7	60.6

Table 6.3 **Generation metric:** Comparison of all generation metrics, JSD, COV, MMD 1-NNA across four categories: Upper Clothing, Lower Clothing, Skirts and Dress/Jumpsuits, for different models. (*Note: Empty entries in the table indicate cases that are not feasible to execute with the previous methods.*)

SMPLicit [CPA*21a] and report quantitative evaluation in Table 6.3. Following recent work [ZVW*22, ADMG18], we use 1-Nearest Neighbor Accuracy (1-NNA) (with both Chamfer distance (CD) and earth mover distance (EMD)) as our main metric. It quantifies the distributional similarity between generated shapes and validation sets and measures quality and diversity [YHH*19]. For a fair comparison, we calculate all metrics using point clouds, rather than meshes. For full comparisons, we also provide other widely used metrics like Jensen-Shannon Divergence (JSD), Coverage (COV), and Minimum Matching Distance (MMD). Prior works, such as DrapeNet, SMPLicit, and ISP, are not trained as generative models but rather as autoencoders. For these methods, we sample a total of 500 samples per class by interpolating in their latent spaces and use these samples to calculate the generation metrics. We use the test set of Cloth3D [BME20] for evaluation. Moreover, only SMPLicit and NeuralABC model body shapes in their latent space, while others [DLLG*23, RGRF22, LGF23] model body shape deformation separately. Hence, we generate both shape-dependent clothing and mean body shape clothing for all the methods and provide shape-dependent clothing evaluation in Table 6.3.

Evaluation metrics for generation experiment

We follow the evaluation metrics used in prior point cloud generation work [YHH*19, ADMG18, ZVW*22] for evaluating the quality of the generation. We define each metric and explain its significance and limitations as a generation metric here.

Coverage (COV) Coverage quantifies the diversity of generated point clouds. It calculates the fraction of reference point clouds matched to at least one generated point cloud using either Chamfer Distance (CD) or Earth Mover's Distance (EMD). We provide both metrics in our evaluation. COV can measure diversity, but it is sensitive to mode dropping and does

not evaluate the overall quality of the generated samples. Diverse but low-quality point clouds can still achieve high coverage scores. We define the COV score using the following equation:

$$\mathrm{COV}(S_g, S_r) = \frac{\left|\{\mathrm{argmin}_{Y \in S_r} D(X,Y) \,|\, X \in S_g\}\right|}{|S_r|} \tag{6.10}$$

where S_g is set of generated point clouds, S_r is set of reference point cloud, and $D(X,Y)$ can be Chamfer Distance or Earth Mover's Distance.

Minimum matching Distance (MMD) MMD measures the overall quality of generated point clouds by calculating the average distance between each reference point cloud and its closest neighbor in the generated set. Like COV, it uses CD or EMD as the distance metric. However, MMD is insensitive to low-quality generated points that are not matched to any reference shapes, making it less reliable for quantifying diversity or coverage. The formula for MMD is given by:

$$\mathrm{MMD}(S_g, S_r) = \frac{1}{|S_r|} \sum_{Y \in S_r} \min_{X \in S_g} D(X,Y) \tag{6.11}$$

1-nearest neighbor accuracy (1-NNA) To overcome the limitations of COV and MMD, 1-NNA evaluates the quality and diversity of generated point clouds. It classifies each point cloud as belonging to either the generated set(S_g) or the reference set (S_r) based on its nearest neighbor, and computes the leave-one-out accuracy of this classification. A perfect match between S_g and S_r results in a 50% classification accuracy, providing a balanced measure of both quality and diversity. 1-NNA is calculated using:

$$\text{1-NNA}(S_g, S_r) = \frac{\sum_{X \in S_g} \mathbb{I}[N_X \in S_g] + \sum_{Y \in S_r} \mathbb{I}[N_Y \in S_r]}{|S_g| + |S_r|} \tag{6.12}$$

where,

- N_X: Nearest neighbor of X in the set $S_g \cup S_r \setminus \{X\}$.
- $\mathbb{I}[\cdot]$: Indicator function.

We use the implementation provided by PVD [ZDW21] to calculate the abovementioned metrics.

Jensen Shanon Divergence (JSD) Jensen-Shannon Divergence (JSD) measures the similarity between the marginal point distributions of the reference set (P_r) and the generated set (P_g). It is computed as the average of the Kullback-Leibler divergence (D_{KL}) of these distributions from their mean distribution (M):

$$\mathrm{JSD}(P_g, P_r) = \frac{1}{2} D_{KL}(P_r \| M) + \frac{1}{2} D_{KL}(P_g \| M), \tag{6.13}$$

	SMPLicit	DIG	DrapeNet	ISP	TailorDiff	GT data
Upper	$14.41^{\pm 11.81}$	$3.68^{\pm 1.83}$	$2.03^{\pm 4.68}$	$1.08^{\pm 5.91}$	$1.43^{\pm 2.95}$	$0.80^{\pm 0.12}$
Lower	$38.09^{\pm 3.791}$	$2.93^{\pm 5.91}$	$3.46^{\pm 8.91}$	$0.69^{\pm 1.08}$	$2.39^{\pm 3.48}$	$0.01^{\pm 1.31}$
Skirt	$24.25^{\pm 4.79}$	-	-	$0.33^{\pm 5.20}$	$1.02^{\pm 1.74}$	$0.11^{\pm 0.36}$
Dress	-	-	-	-	$2.05^{\pm 3.91}$	$0.02^{\pm 0.29}$

Table 6.4 **Body-Cloth Intersection:** We compare the body-cloth intersection percentage (% ↓) across different models and clothing categories.

where:

$$M = \frac{1}{2}(P_r + P_g)$$

and,

- P_r and P_g are the marginal distributions of points in the reference and generated sets, respectively.
- These distributions are estimated by discretizing the space into 28^3 voxels and assigning each point to one voxel.

While JSD effectively captures differences in marginal distributions, it ignores the structural relationships between individual shapes. Consequently, a model that always produces the "average shape" may achieve a perfect JSD score, despite failing to learn meaningful variations in the data. For JSD, we use the implementation provided by PointFlow [YHH*19].

We provide the quantitative evaluation across all clothing classes using all methods in Table 6.3. For DrapeNet [DLLG*23], we filter out the results where draping fails. As pointed out by previous works [ZDW21, ZVW*22], COV and MMD are not very reliable metrics as they do not assess the quality of the generated results. We also observe more inconclusive values for these metrics. For instance, SMPLicit [CPA*21a] tends to have significantly low MMD scores for skirts and lower clothing; however, visually, SMPLicit often results in holes in the extracted meshes, which increases the diversity of the generated samples but does not necessarily indicate that the generated samples are of good quality.

We further evaluate the quality of generation using the body-cloth intersection percentage. Specifically, we calculate the percentage of clothing points located inside the body mesh. For all methods, we use the same body shape parameters and generate clothing like that introduced above. We then sample a fixed number of points from the mesh surface/point cloud and calculate the percentage of points inside the body. We report the average percentage across different clothing categories in Table 6.4. Although ISP [LGF23] generates the lowest amount of intersection, it is worth noting that it exhibits very limited diversity, as seen in Table 6.3. This is also evident in the clothing reconstruction experiment (see Table 6.5), where ISP struggles to effectively represent unseen clothing.

6.4.2 Cloth Editing using TailorDiff

In this section, we demonstrate how TailorDiff can be effectively utilized to modify clothing styles by traversing its learned latent space. The model offers a wide range of controllability, enabling both high-level structural changes and subtle, fine-grained modifications. In the subsequent sections, we delve deeper into how these varying levels of control can be leveraged to perform diverse and precise clothing edits. Additionally, we discuss the limitations of existing methods, emphasizing their fragility and shortcomings in supporting downstream applications.

6.4.2.1 Cloth Style editing

To showcase clothing style editing for a fixed body shape, we present qualitative results in Figure 6.7. In particular, we modify the style of the input clothing by performing interpolation within the latent space of the model. We also provide a comparison with prior works, which tend to be restricted to specific clothing categories and are unable to capture the full diversity of garments generated by TailorDiff. For a fair comparison, we randomly selected a representative prior work corresponding to each clothing category under consideration. When available, we also include results for additional clothing items.

Prior works typically offer a list of predefined latent codes corresponding to their models. For each such model, we manually selected the latent code that most closely matches the clothing shape produced by TailorDiff. For example, NeuralABC [HYZ24], which is limited to generating combined clothing types, was evaluated using a dress example (Figure 6.7, top). However, NeuralABC often suffers from surface inconsistencies such as holes, a limitation we attribute to its underlying meshUDF [GSF22] representation. Similar issues arise with SMPLicit, which also relies on a UDF-based representation (Figure 6.7, middle). In the pants category, SMPLicit exhibits visible geometry artifacts, especially near seams, indicating challenges in maintaining surface consistency. These artifacts can be largely attributed to both the restricted diversity in the training dataset and the less structured nature of the model architecture. For skirts, we compare against ISP [LGF23], a method known for producing high-quality meshes. Despite its quality, ISP remains a category-specific model and does not generalize well across clothing types. Moreover, its capacity for generating diverse garments is limited.

In contrast, TailorDiff is a single, unified model capable of generating a wide variety of clothing, offering flexibility and fine-grained control across all clothing categories. Since TailorDiff is trained on a large-scale dataset and has a hierarchical latent space, it is highly expressive, providing fine-grained control. Its continuous latent space supports meaningful interpolations and nuanced edits without retraining or category-specific tuning. To further illustrate its capabilities, we provide additional results for coarse-level editing for Dresses in Figure 6.4. Here, we generate two random clothing samples for a fixed body shape and interpolate between their respective latent representations. The resulting interpolated outputs are consistently valid, demonstrating that the latent space of TailorDiff is smooth and well-behaved. These interpolations yield novel garment designs, exhibiting seamless

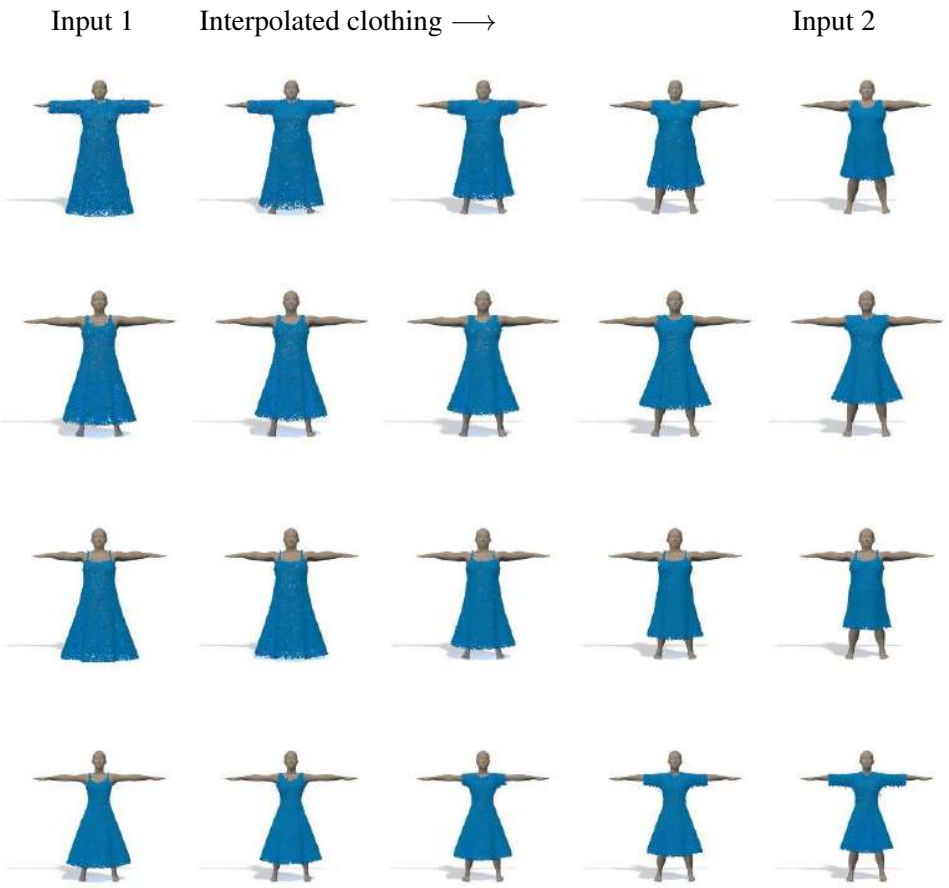


Fig. 6.4 **Cloth style editing by interpolation for Dress:** Given two input clothing, we show editing by interpolating between the latents.

transitions in features such as sleeve length, neckline shape, and overall garment structure. We provide more results for Skirts in Figure 6.5 and for Jumpsuits in Figure 6.6.

6.4.2.2 Fine-grain Editing

TailorDiff also has the ability for fine-grained editing, similar to [ZVW*22]. For a given clothing shape, TailorDiff enables the generation of controlled variations, offering a flexible tool for detailed design exploration. This is done through two distinct approaches.

First, we use a *diffuse-denoise* based strategy. Given a clothing point cloud $\mathbf{x}$, we encode it into two latent representations: the global shape latent $\mathbf{z}_0$ and the local detail latent $\mathbf{h}_0$. These latents are then subjected to a forward diffusion process for a small number of time steps, $\tau < T$, introducing slight stochastic noise. Subsequently, we apply the reverse diffusion process to denoise the latent representation and reconstruct a variant of the original clothing shape. This results in new samples that retain the core structure while exhibiting subtle variations in fine-grained details. See Figure 6.8(top), where given a sleeveless dress, our diffuse-denoise approach adds small details to the silhouette and hemline of the dress. Such

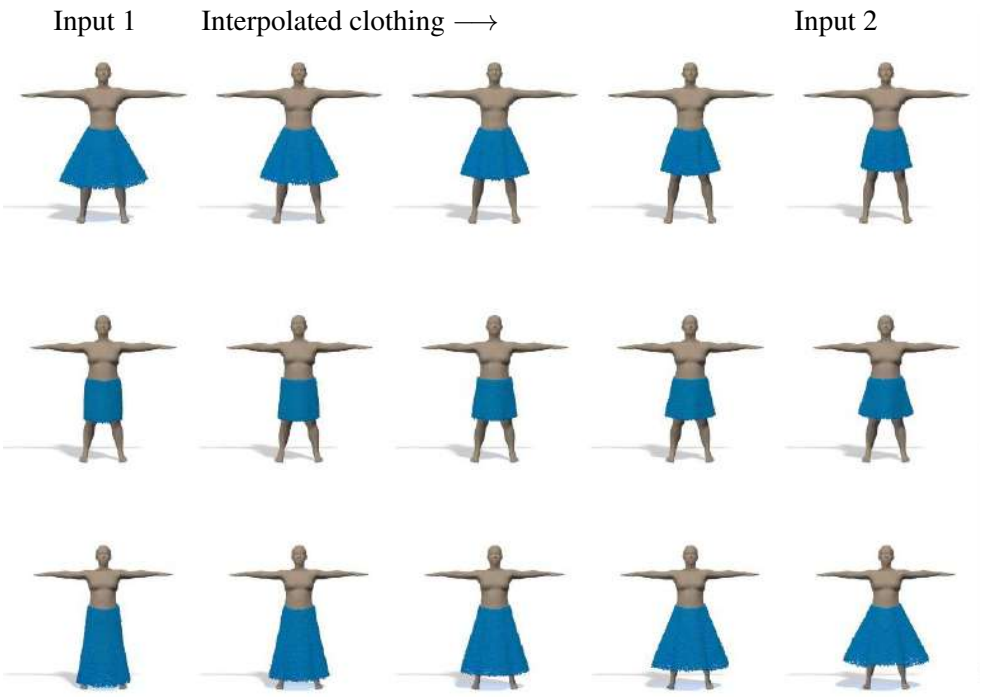


Fig. 6.5 **Cloth style editing by interpolation for Skirts:** Given two input clothing, we show editing by interpolating between the latents.

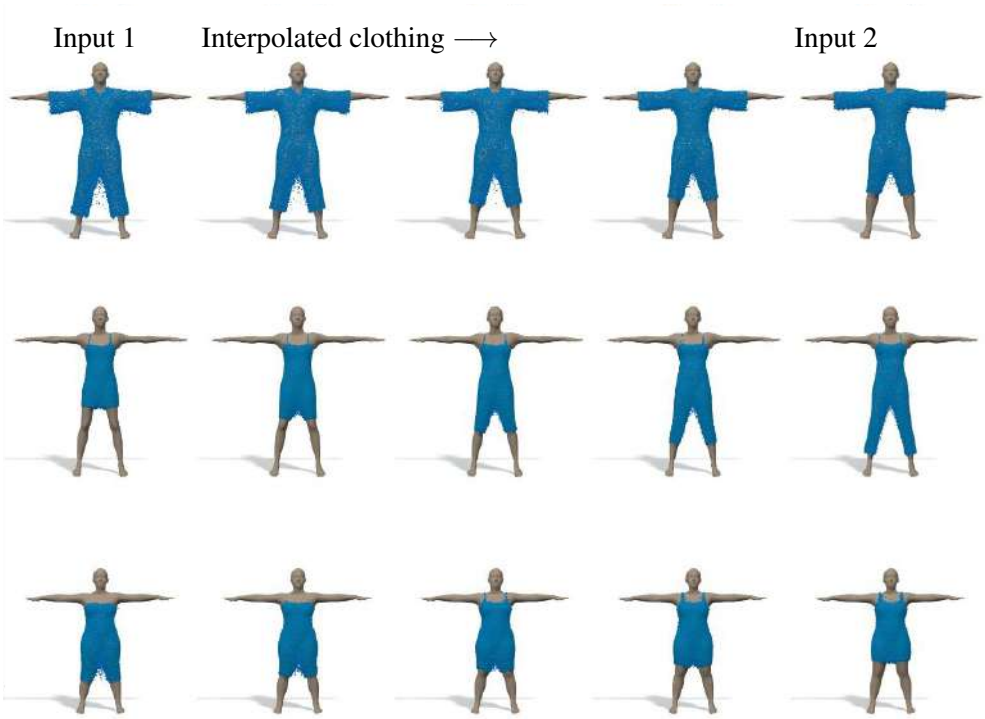


Fig. 6.6 **Cloth style editing by interpolation for Jumpsuits:** Given two input clothing, we show editing by interpolating between the latents.

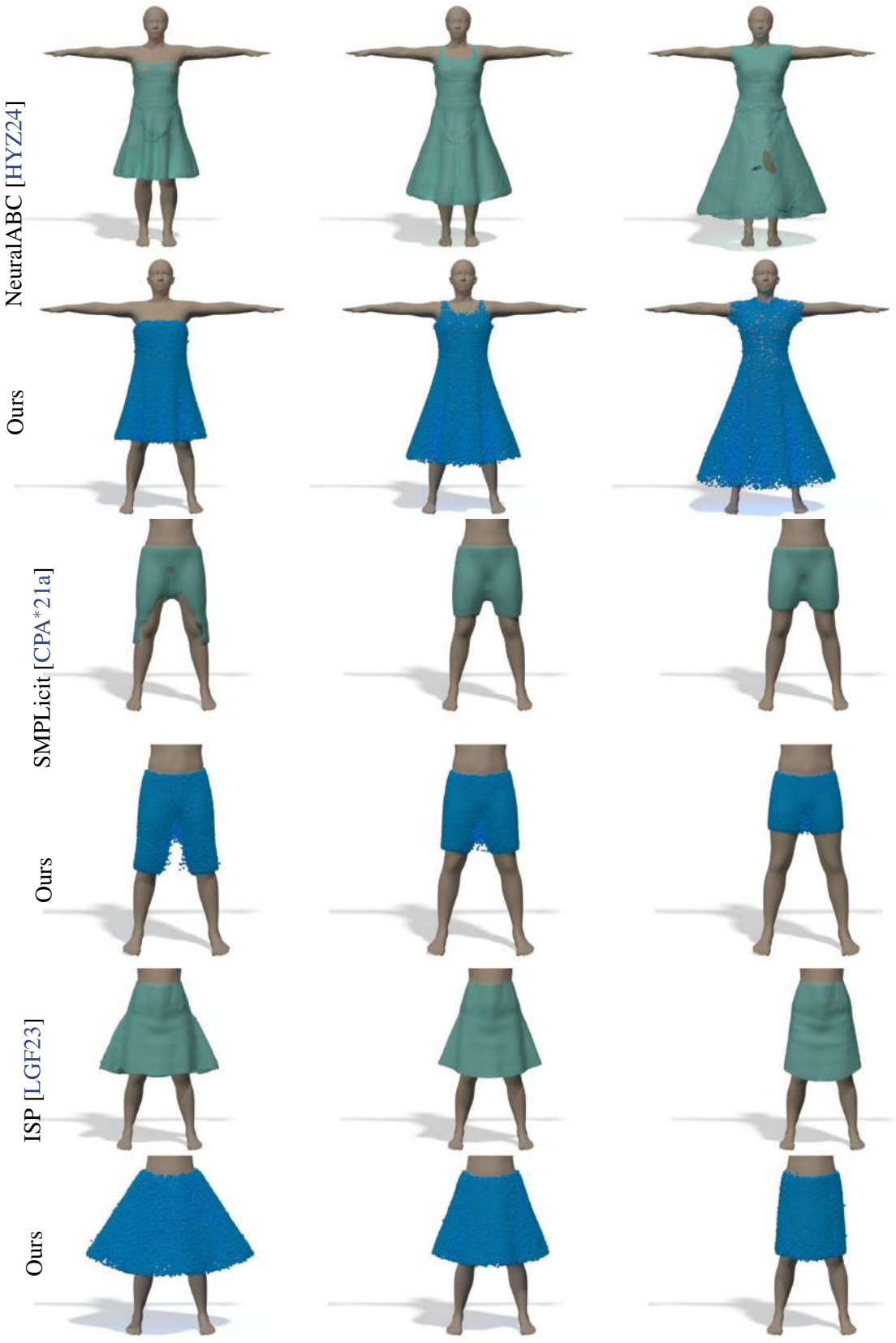


Fig. 6.7 **Comparison with prior work for cloth style editing:** For a fixed body shape, we edit the clothing style by interpolating in latent space. Specifically, the middle column is generated by blending the latent codes corresponding to the clothing styles shown in the left and right columns.

a method is particularly advantageous for applications in digital fashion design and virtual prototyping, where designers may wish to explore stylistic variations while preserving the overall garment design.

Second, TailorDiff allows for explicit control over the *clothing occupancy or coverage on the body*, denoted as **o**. By adjusting this occupancy representation, encoded as $\mathbf{z_c}$, users can directly manipulate interpretable design elements such as sleeve length, neck shape, garment length, and other structural features. As shown in Figure 6.8 (bottom), this method enables precise modifications to specific regions of the garment, such as altering the neck style from round to V-shaped to strapless. In Figure 6.9, we further showcase a variety of fine-grained edits across diverse clothing types. We show examples of increasing the length of the dress, skirt, t-shirt *etc.*, and increasing the volume of the dress. The results demonstrate that TailorDiff effectively performs the desired fine-grained edits while maintaining other controls fixed. This capability is particularly valuable for clothing design, allowing precise adjustments without compromising other attributes.

Overall, these capabilities make TailorDiff a powerful and intuitive tool for generating controllable clothing. It supports both unconstrained design exploration through stochastic perturbations and targeted edits through explicit control variables. This dual functionality makes it highly valuable for use cases in computer-aided fashion design, virtual try-on systems, and the generation of customizable avatars.



Fig. 6.8 **Fine-grained editing:** Given an input clothing(*left*), we show fine grain-control on the clothing generated. *Top*: We use diffuse-denoise to modify the clothing style, and *Bottom*: We use **o** to control the sleeve style, while keeping other controls fixed.

As seen in Figure 6.10, prior works frequently exhibit significant geometric artefacts, particularly when generalizing across diverse body shapes and clothing styles. For instance, DrapeNet [DLLG*23] often produces unrealistic draping results for certain combinations of latent variables and body shapes. This is primarily due to its self-supervised draping

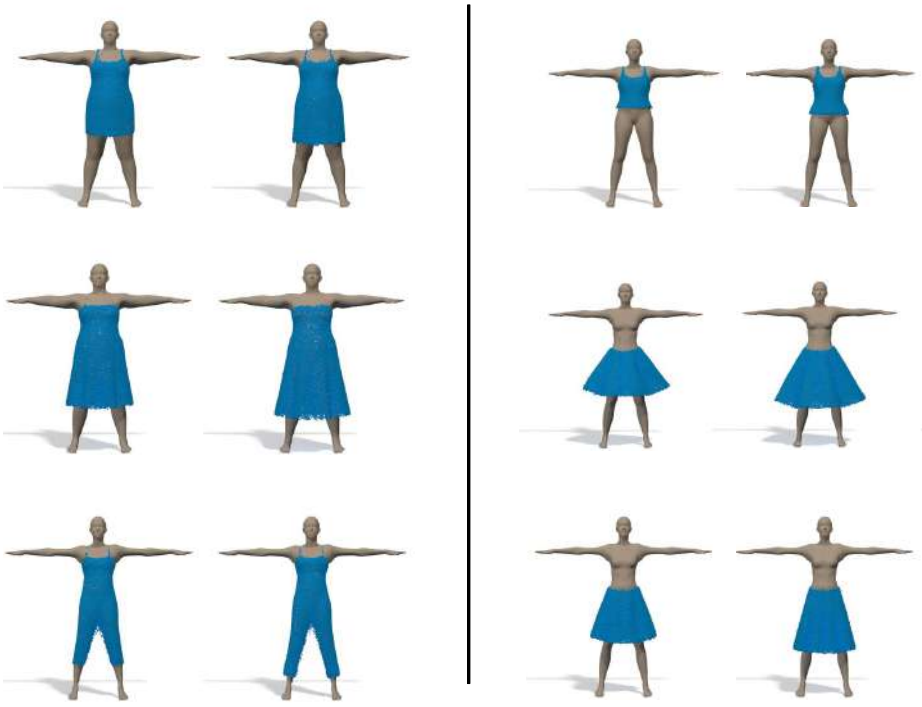


Fig. 6.9 **More fine-grained editing examples for diverse clothing styles:** Given an input clothing(*left*), we show fine-grained control on the clothing generated. We only change the length here by adjusting the cloth occupancy on the body, while keeping all other controls fixed.

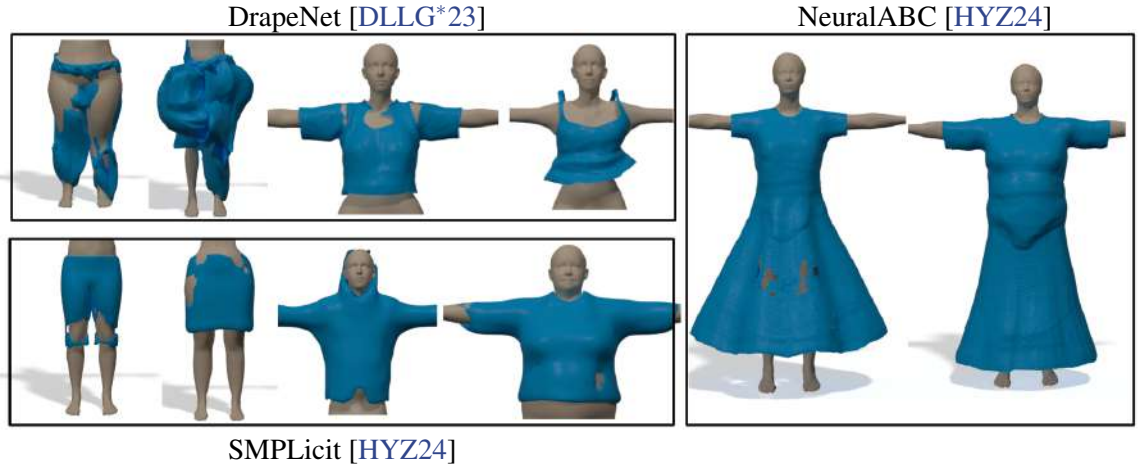


Fig. 6.10 **Failure cases of prior clothing models:** We observe that prior work such as DrapeNet [DLLG*23] often results in unrealistic draping for new shape of body or clothing. SMPLicit [CPA*21a] often results in missing geomtery and big holes. NeuralABC [HYZ24] results in holes and bumpy artefacts near torse.

framework, which lacks strong constraints to ensure physical plausibility and generalization to unseen body-cloth configurations. Furthermore, SMPLicit [CPA*21a] also suffers from several notable limitations. In multiple examples, it fails to generate complete clothing geometries, as evidenced by missing parts in garments such as pants, shirts, and hoodies. Additionally, SMPLicit exhibits visible body-cloth interpenetrations, particularly in garments like skirts, where the clothing meshes intersect undesirably with the underlying body. These shortcomings can be largely attributed to the limited diversity of the training dataset used in SMPLicit, which hinders the model's ability to generalize across different garment categories and body types.

In contrast, interpolations within our TailorDiff model exhibit strong semantic coherence, maintaining consistent structural details throughout the interpolation path. We will see more detailed results in Section 6.4.3. This highlights the robustness and expressiveness of our learned clothing manifold, which is capable of modeling complex intra-class variations without sacrificing geometric fidelity or garment-body alignment. Such robustness is crucial for downstream applications, making TailorDiff a powerful and reliable tool for a wide range of tasks, including scan fitting, image fitting, content generation *etc.*.

6.4.3 Body Shape Editing using TailorDiff

Similar to clothing style editing, we also enable editing of the underlying body shape, which allows for the generation of clothing that is not only visually plausible but also explicitly compatible with the modified body geometry. This functionality is essential for achieving realistic and personalized virtual representations. We demonstrate the effectiveness and flexibility of our approach on a diverse and challenging set of body shapes in Figure 6.11, showcasing how our model adapts seamlessly to different physical proportions. We observe that TailorDiff is capable of accurately modeling even the most extreme and atypical body shapes, maintaining high fidelity in both the body representation and the corresponding clothing geometry. This ability is further shown in the rightmost column of Figure 6.1, where our model successfully handles significant deviations from average body proportions while preserving garment structure and alignment.

In TailorDiff, a key contribution lies in the explicit incorporation of body encoding, which plays a crucial role in significantly enhancing the alignment between the generated clothing and the underlying body shape. By conditioning the model on detailed body information, our approach ensures that garments conform closely to the body's surface, thereby avoiding unrealistic intersections or ill-fitting results. Our proposed body-aware encoding is designed to generalize seamlessly across a wide range of body types, including variations in gender, body mass, and proportion. This generalization capability allows TailorDiff to maintain robust performance even in the presence of highly challenging or uncommon body shapes, consistently producing well-fitted clothing that respects both anatomical accuracy and garment structure. Importantly, the generated clothing remains free from collisions or penetrations with the underlying body, which is a common failure case in existing approaches.

In this section, we present a comparison between TailorDiff and several state-of-the-art clothing generation methods [CPA*21a, LGF23], with a specific focus on the task of body

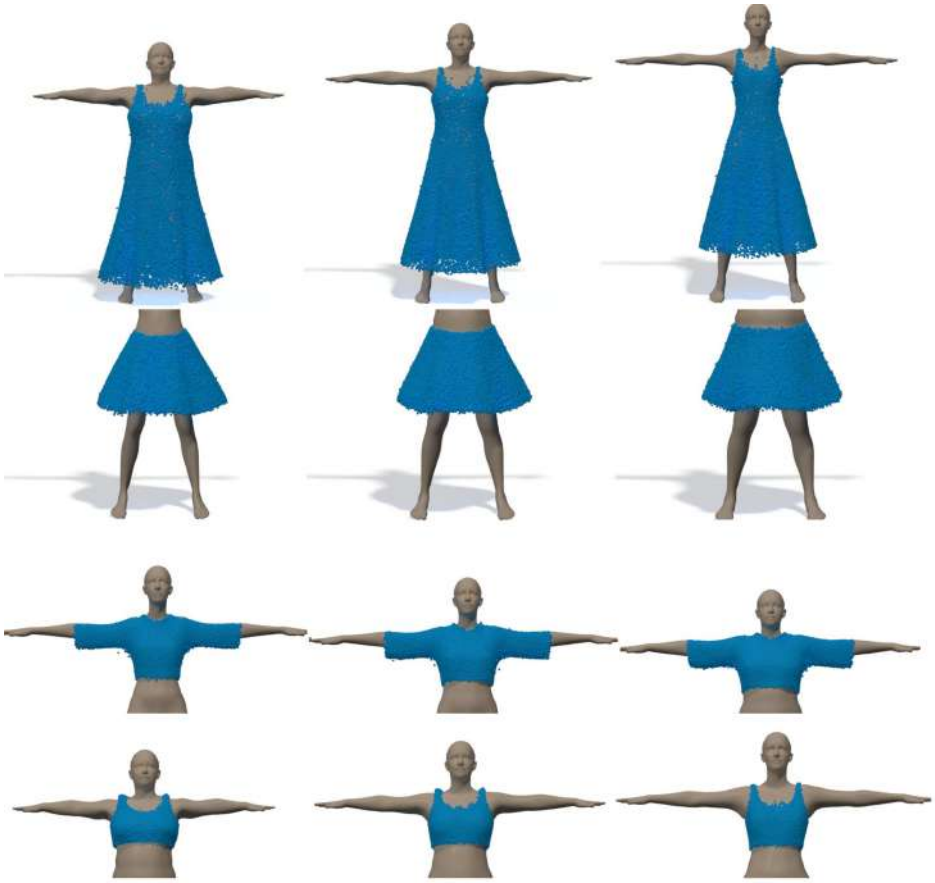


Fig. 6.11 **Body Shape Editing:** For a given clothing, we change the underlying body shape. TailorDiff adapts perfectly to very diverse body shapes.

shape editing. Through this comparison, we demonstrate that TailorDiff not only performs reliably across different body shapes but also sets a new benchmark in producing visually coherent and physically plausible clothing in shape-conditioned generation scenarios.

We compare our approach only with SMPLicit [CPA*21a] and NeuralABC [HYZ24], as these are among the few existing models that explicitly condition body shape during clothing generation. These methods take the underlying body geometry into account in a direct and meaningful way when synthesizing clothing. In contrast, other methods such as DrapeNet [DLLG*23], ISP [LGF23], and DIG [RGRF22] primarily apply body-shape-dependent deformations as a post-processing step, where pre-generated clothing is draped onto new body shapes.

We visualize our comparison with SMPLicit [CPA*21a] in Figure 6.12. The qualitative results clearly demonstrate that SMPLicit often struggles to generate faithful and complete clothing geometry, particularly when the underlying body shape tends toward being overweight or deviates significantly from average body shape. These limitations manifest as distortions, missing parts, or strong intersections between body and clothing. For NeuralABC [HYZ24], we provide results in Figure 6.13. We provide results for NeuralABC in the

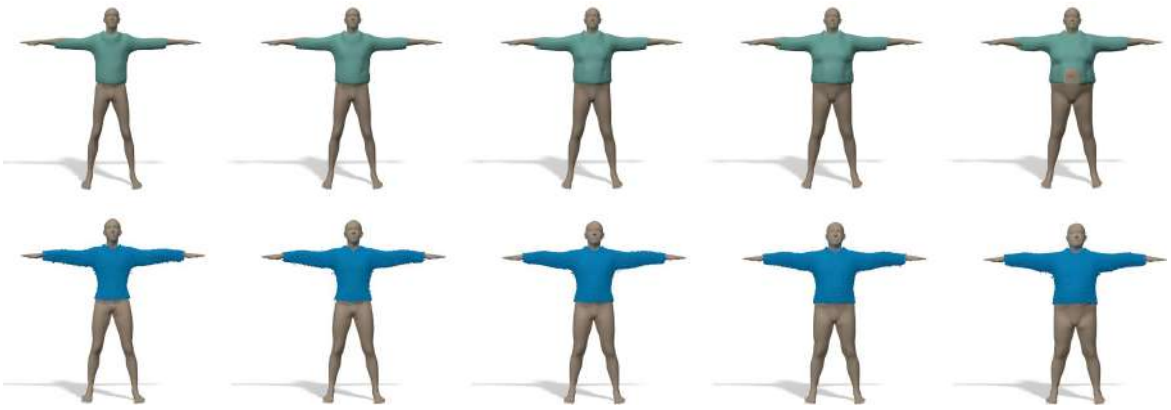


Fig. 6.12 **Body shape editing comparison with SMPLicit [CPA*21a].** As shown in the figure, SMPLicit struggles to generate a well-fitting shirt and exhibits collisions with the underlying body, sometimes resulting in holes in the clothing mesh. In contrast, our model adapts effectively to diverse body shapes, producing high-quality results without collisions.

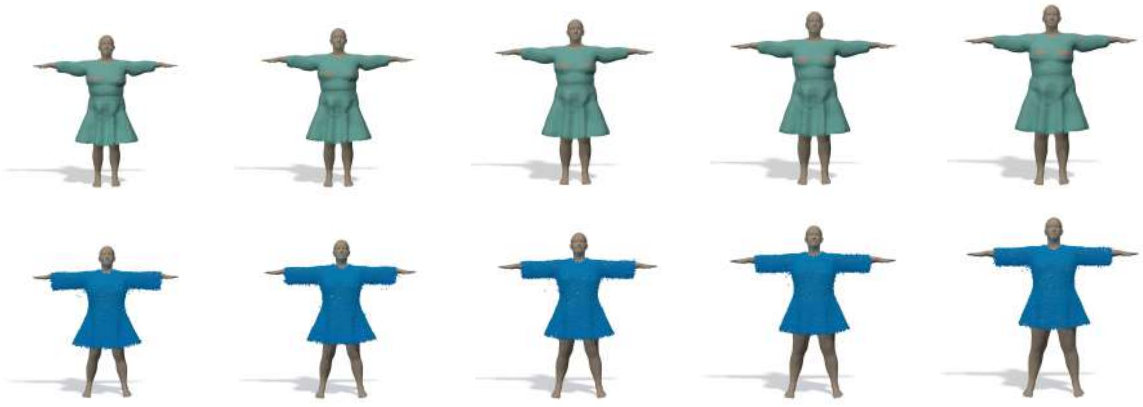


Fig. 6.13 **Body shape editing comparison with NeuralABC [HYZ24].** As shown in the figure, clothing generated by NeuralABC exhibit collisions with the underlying body, particularly in the chest region. In contrast, our TailorDiff avoids such issues, consistently producing well-fitted clothing for diverse body shapes.

T pose, which aligns with their modeling setup, as they generate clothing in a canonical T pose. In contrast, our results are shown using the TA pose, which introduces articulation and better reflects realistic use cases. Despite this added complexity, our TailorDiff consistently generates high-quality, well-fitted clothing without visible collisions or artefacts, thanks to its novel body-aware architectural design. Moreover, unlike prior work that often requires separate handling for different genders, our model is inherently gender-agnostic. This enables it to generalize more effectively across diverse identities, resulting in significantly improved body-cloth layering and contributing to the overall realism and robustness of our clothing generation pipeline.

We present additional shape editing results across a wide variety of clothing styles in Figure 6.14 and Figure 6.15, further highlighting the versatility and robustness of our approach. These figures showcase the ability of TailorDiff to handle diverse and challenging body shapes while maintaining the semantic integrity and structural consistency of the garments.

The results clearly demonstrate that our TailorDiff excels in the task of shape editing, consistently producing high-quality and visually coherent outputs. Importantly, this holds true across a broad range of clothing categories, including garments with complex structure and varying levels of articulation. Our model handles jumpsuits, skirts, dresses, tops, and T-shirts with equal ease, generating garments that not only adapt to the underlying body shape but also preserve their characteristic design features and fit naturally on the modified body. This further underlines the effectiveness of our body-aware representation in enabling generalizable and realistic clothing generation across different styles and shapes.

6.4.4 Clothing Reconstruction

We further evaluate the capability of our model and its latent space in the clothing reconstruction task, comparing it with prior works that were also evaluated on the same task. Specifically, for encoder-decoder-based architectures, we encode and decode unseen clothing, reporting the reconstruction error in Table 6.5. For DrapeNet and WordRobe, we report numbers from WordRobe [SMR*24]. For ISP [LGF23], we use the mesh fitting code provided in their work, as it is an auto-decoding-based method.

In addition to the quantitative comparisons of clothing reconstruction provided in the main paper, we include further qualitative comparisons with ISP [LGF23]. Reconstruction results for skirts and shirts are shown in Figure 6.16. The results clearly illustrate that ISP struggles to accurately reconstruct target clothing, both in terms of clothing style and underlying body fit. We attribute this limitation to the auto-decoding latent space used by ISP, which lacks the robustness and capacity to handle diverse clothing styles. In contrast, our model leverages a generative model in the latent space, which inherently enables superior clothing modeling capabilities and results in significantly better clothing reconstruction performance.

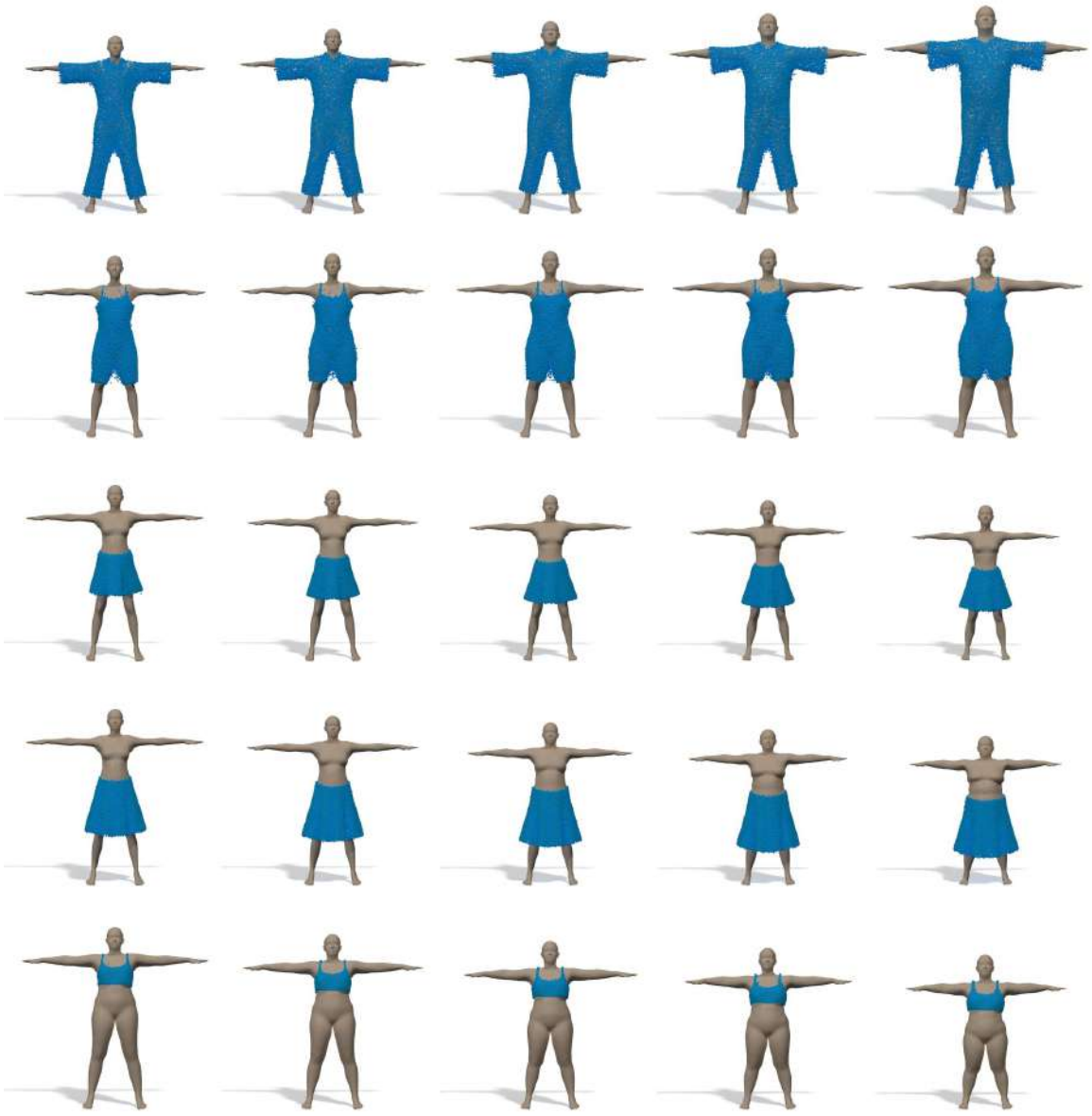


Fig. 6.14 **Body shape editing for more different clothing styles:** Our body-aware generation mechanism is capable of adapting to diverse and challenging body shapes, whether short, tall, slim, or overweight. The generated clothing aligns seamlessly with the underlying body and does not result in collisions.

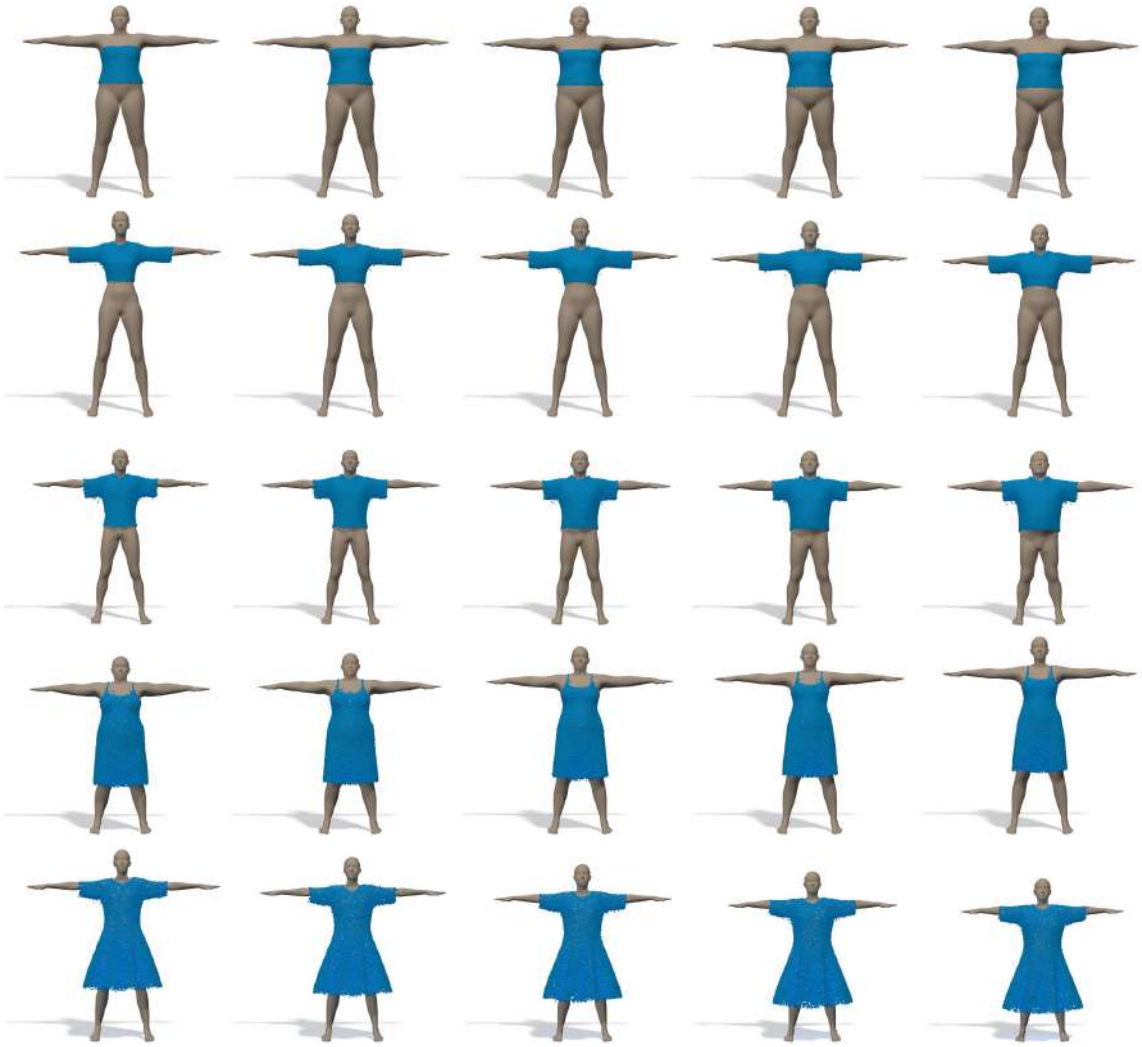


Fig. 6.15 **Body shape editing for more different clothing styles:** Our body-aware generation mechanism is capable of adapting to diverse and challenging body shapes, whether short, tall, slim, or overweight. The generated clothing aligns seamlessly with the underlying body and does not result in collisions.

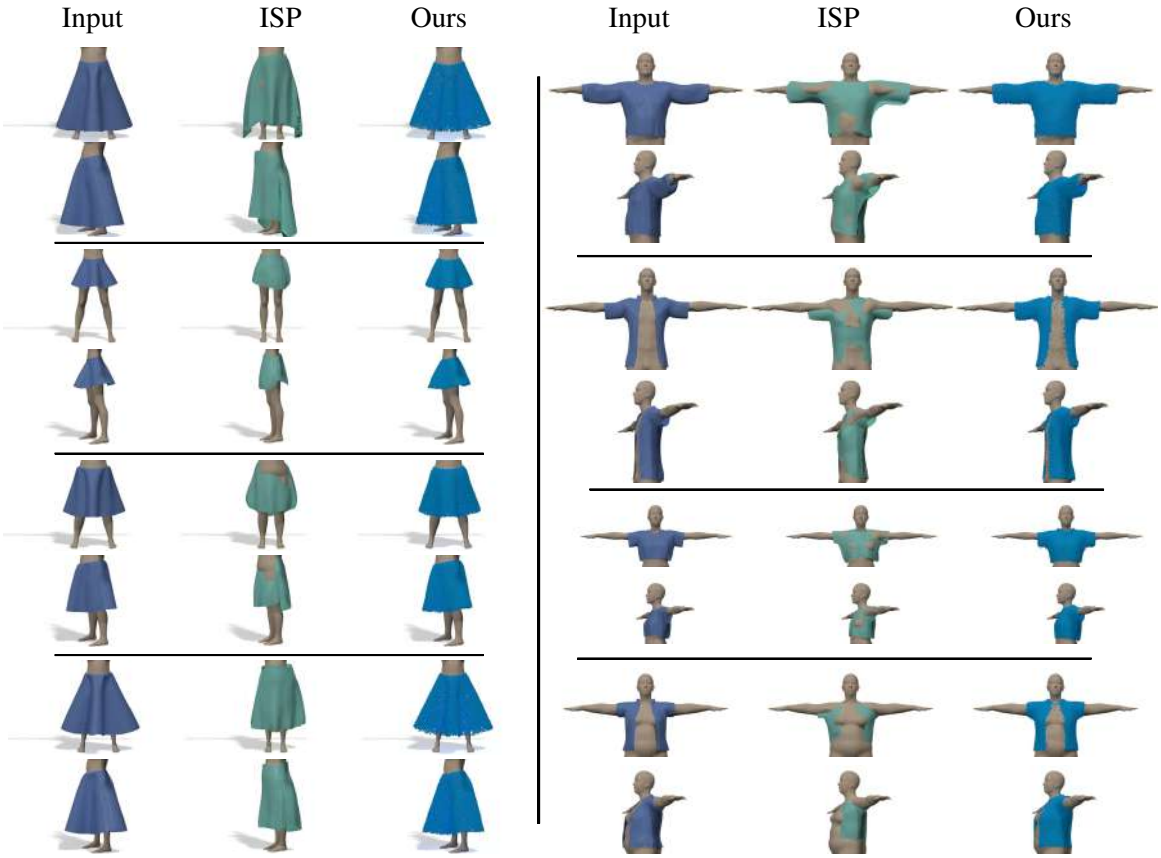


Fig. 6.16 **Clothing reconstruction of Skirts and Shirts:** We visualize the target 3D clothing (*left*), the reconstruction by ISP [LGF23] (*middle*), and the reconstruction by our method (*right*). ISP struggles to accurately reconstruct the clothing, resulting in collisions with the body. In contrast, our body-aware mechanism effectively avoids such limitations.

Method	Upper Clothing		Pants		Skirt	
	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓
DrapeNet [DLLG*23]	1.522	0.631	1.585	0.739	-	-
ISP [LGF23]	2.16	1.16	2.66	2.47	3.05	1.63
WordRobe [SMR*24]	1.491	0.635	1.568	0.703	-	-
Ours	**0.826**	**0.549**	**0.846**	**0.599**	**1.249**	**0.956**

Table 6.5 **Clothing reconstruction:** Evaluation results of clothing reconstruction on CLOTH3D test set (*in cm*).

6.4.5 Fitting TailorDiff to Scans

In this experiment, we demonstrate the capability of TailorDiff to fit 3D scans of humans with clothing segmentation. Given a scan $\mathcal{S}$, with SMPL parameters θ, β and clothing segmentation labels, we devise an optimization-based method, where we optimize the latent code of TailorDiff to fit the scan observation. We initialize $\mathbf{z}_0$ and $\mathbf{h}_0$ by encoding the segmented clothing from the scan and $\mathbf{z_c}$ by encoding the occupancy of each clothing in the SMPL body. This provides already good enough initialization for latent variables. We then optimize $\mathbf{z_c}$, $\mathbf{z}_0$, and $\mathbf{h}_0$ to minimize both reconstruction loss and an intersection loss, which ensures that clothing stays outside the body. We use TailorDiff to get a layered representation from single-layered scans with per-point clothing segmentation labels. The challenge of fitting a model to segmented scans is that inner clothing is occluded and this results in a significant loss of shape information, see Figure 6.19(top). For this, we devise an optimization method, which optimizes for the latent variables of TailorDiff and minimizes the loss given in Equation (6.14). For the sake of brevity, we use v_i to represent all the latent codes of TailorDiff. Specifically, $v = \{\mathbf{z}, \mathbf{h}, \mathbf{z_c}\}$ and i is the clothing layer id, such that $i \in [0, l)$, where l is the number of clothing layers, including body, and we assign a lower layer id to inner layers. $\mathcal{S}$ represents the target scan with segmentation labels, such that$\mathcal{S}_i$ is the submesh which represents clothing layer i. $P(\theta, \beta, v_i)$ represents the posed point cloud using SMPL pose and shape parameters (θ, β), and v_i is used to get a canonical point cloud from TailorDiff. $\mathcal{D}_{\mathcal{S}_i \rightarrow P(\theta,\beta,v_i)}$ is the unidirectional chamfer distance between segmented clothing($\mathcal{S}_i$) and the posed clothing point cloud. We only use unidirectional distance in this case as the whole surface of the clothing layer might not be visible (see Figure 6.19). We also minimize the overall bidirectional chamfer distance between a pointcloud obtained by accumulating all the layers, which is given by $P(\theta, \beta, v_0, \ldots, v_l)$ and $\mathcal{S}$. This helps the occluded part of clothing $P(\theta, \beta, v_i)$ to get some information about the hidden shape. We further introduce a layer-preserving loss, which ensures that the inner layer of clothing stays inside the outer layer. Specifically, we penalise if the distance of the inner layer from the

SMPL mesh is more than the distance of the outer layer from the SMPL mesh.

$$\begin{aligned}\mathcal{L}(v_0,\ldots,v_l) = & \mathrm{w}_{\text{data}} \sum_{i=0}^{l} \mathcal{D}_{\mathcal{S}_i \to P(\theta,\beta,v_i)}(\mathcal{S}_i, P(\theta,\beta,v_i)) \\ & + \mathcal{D}(\mathcal{S}, P(\theta,\beta,v_0,\ldots,v_l)) \\ & + \sum_{i=0}^{l} \sum_{j=0}^{i-1} \max(0, d_j - d_i) \\ & + \mathrm{w}_\beta \|\beta - \beta_{\text{init}}\| + \mathrm{w}_\theta \|\theta - \theta_{\text{init}}\|\end{aligned} \tag{6.14}$$

We initialize the v_i by encoding the segmented clothing using the VAE of TailorDiff. This gives a good initialization. We also refine body pose and shape parameters, but keep them close to the initialization. We present the results of our optimization method on the Thuman2.0 dataset, as it comprises diverse and complicated clothing in various poses.

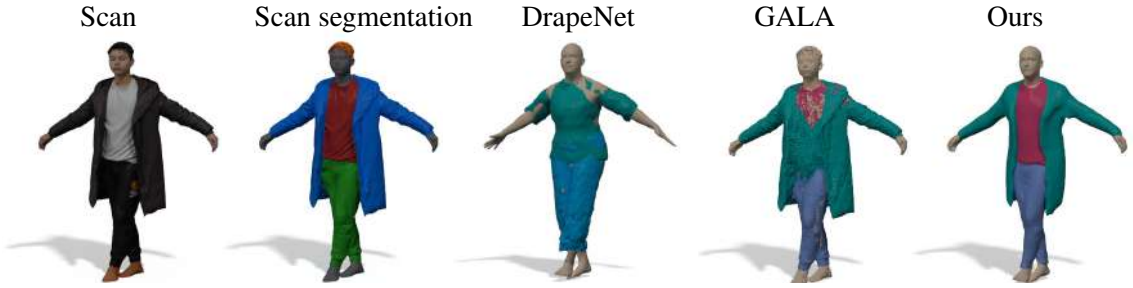


Fig. 6.17 **Layered scan fitting comparison with prior works:** We compare layered scan fitting with GALA [KKSJ24] and DrapeNet [DLLG*23]. DrapeNet struggles to model 3-layer clothing. GALA creates volumetric meshes that do not reflect the reality of the clothing surface, and also result in significant intersections.

Since no unified model currently exists, none of the prior works can fit scans with diverse clothing. Moreover, for prior works like DIG, DrapeNet, ISP, SMPLicit, and NeuralABC, the code for scan fitting is not publicly available. We implemented a DrapeNet-based scan fitting method, following the approach outlined in the paper. We compare our scan fitting results with DrapeNet (as shown in the main paper) and also with GALA [KKSJ24].

GALA is a layered asset generation method that extracts clothing layers from scans sequentially. Specifically, GALA relies on SAM for clothing segmentation labels and uses SDS loss on the DMTet [SGY*21] representation to extract layers one by one. This process takes approximately 30–40 minutes per clothing layer. We demonstrate the results on a complex scan with three clothing layers in Figure 6.17. For better visualization of layers, we convert our fitted point cloud to meshes before rendering. We use a TSDF fusion-based method to obtain meshes from pointcloud [SG18]. The smooth reconstructed mesh also shows that the quality of the point cloud generated by TailorDiff is good and usable for downstream mesh-driven tasks. DrapeNet-based method struggles to represent surfaces with three layers of clothing effectively. For GALA, we observe that the coat incorrectly explains a significant portion of the T-shirt(see Figure 6.17 for the intersection between t-shirt and

coat). GALA is sensitive to SAM segmentation, and since it uses DMTets to represent clothing, the extracted layers are not thin surfaces but rather voluminous meshes, which leads to intersections between layers.

We provide a detailed comparison of GALA and our method in Figure 6.19. GALA produces noisy surfaces for occluded layers, whereas our method generates smooth meshes that accurately capture the geometry of both visible and occluded clothing layers. This is attributed to our strong geometry prior model. We show more results in Figure 6.18.

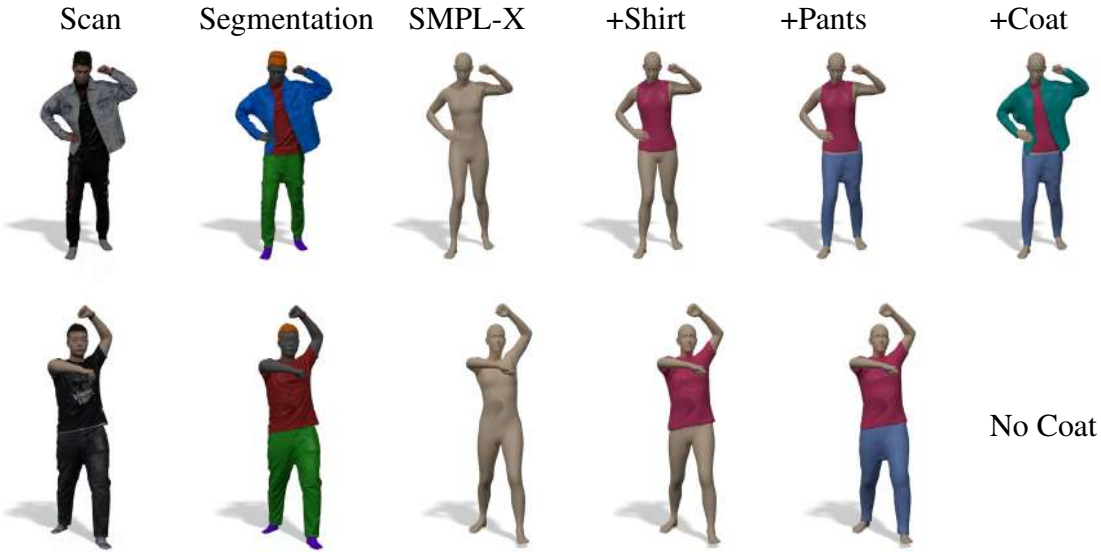


Fig. 6.18 **Layered scan fitting on more scans with diverse clothing:** TailorDiff based fitting enables layer-wise fitting across a wide variety of clothing types, something that prior works cannot achieve.

6.4.6 Ablation

We conduct ablation studies to assess the contribution and significance of the key components of our proposed method, TailorDiff. In particular, we focus on evaluating the importance of our novel body-aware encoding mechanism, which is designed to capture the local geometric context of the clothing in relation to the underlying human body. To this end, we train a variant of our model in which the body-aware encoding is entirely removed. In this ablated configuration, we replace our custom Body-PVConv layers with standard PVConv layers that do not explicitly incorporate body information. This allows us to isolate and measure the specific impact of body-aware features on the performance of our method. Furthermore, we investigate the role of the point latent space in our representation. In this setup, we ablate the use of per-point latent features and instead condition the model solely on a global latent code, denoted as $\mathbf{z}$, which encapsulates the overall shape information without capturing localized details. This experiment enables us to examine how critical the spatially distributed latent encoding is to the expressiveness and accuracy of the learned clothing representations.

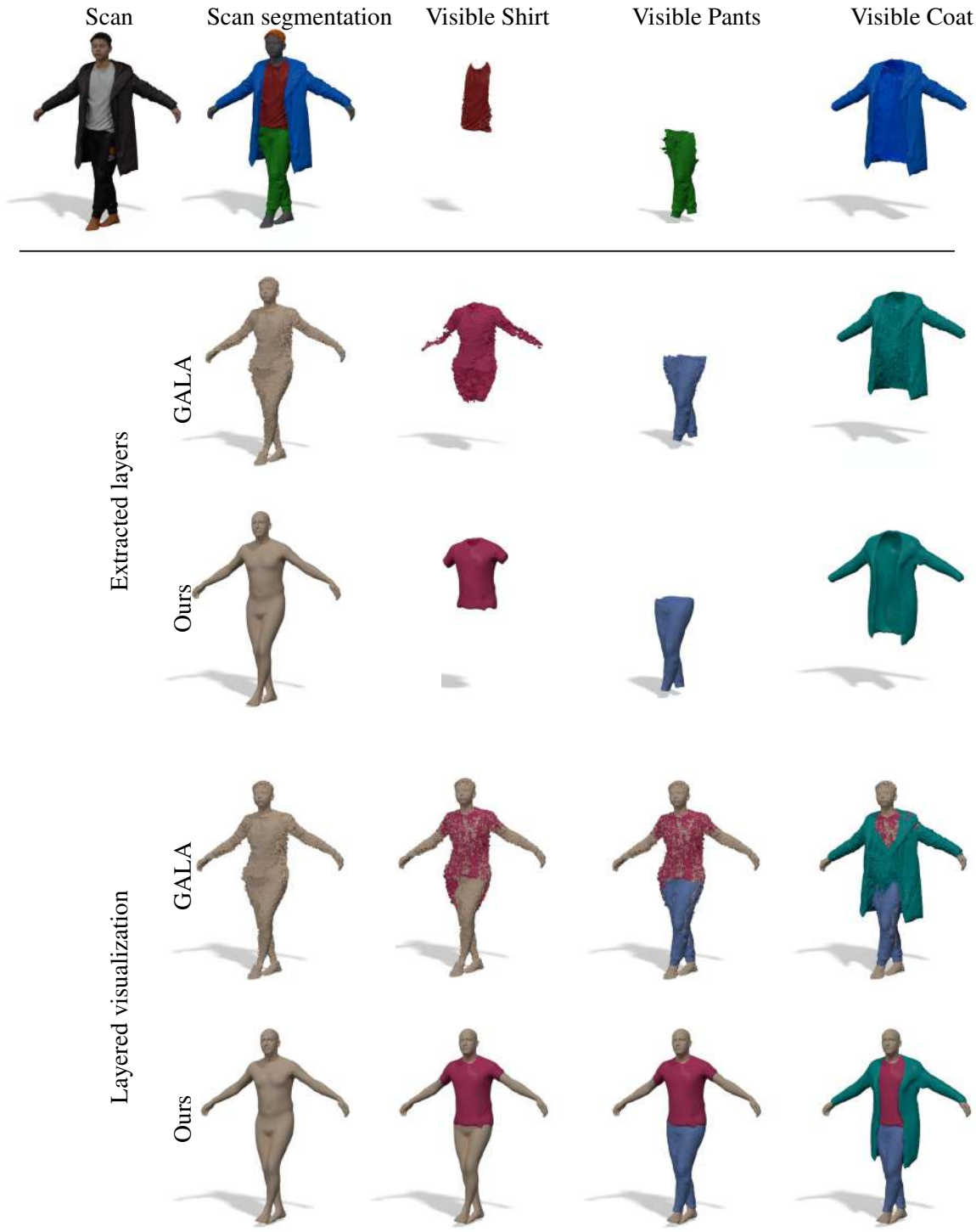


Fig. 6.19 **Layered clothing assets from scans:** Given an input scan with segmentation labels(*top*), where only partial surfaces of the inner clothing are visible, our scan fitting method effectively fits multiple clothing layers to the scan while preserving the style and structure of the clothing. In contrast, GALA [KKSJ24] can achieve similar results but produces noisy meshes with incorrect shapes (*see Extracted layers*) and significant intersections between layers (*see Layered visualization*). Our method, however, generates accurate shapes and styles of clothing with proper layer ordering, ensuring no intersections between layers.

Metric	Global only	w/o Body enc.	Full Model	Full w/ inter. loss	Train. set
1-NNA-CD↓	71.5	68.8	64.6	**64.5**	59.7
1-NNA-EMD↓	70.6	67.6	**59.7**	61.7	56.1
Intersection ↓	$8.83^{\pm 5.27}$	$5.47^{\pm 8.90}$	$\mathbf{1.73}^{\pm \mathbf{4.71}}$	$\mathbf{1.73}^{\pm \mathbf{3.84}}$	$0.24^{\pm 0.56}$

Table 6.6 **Ablation on network architecture and latent features:** We evaluate the impact of body encoding and a hierarchical latent space on diversity (1-NNA %) and realism (body-cloth intersection).

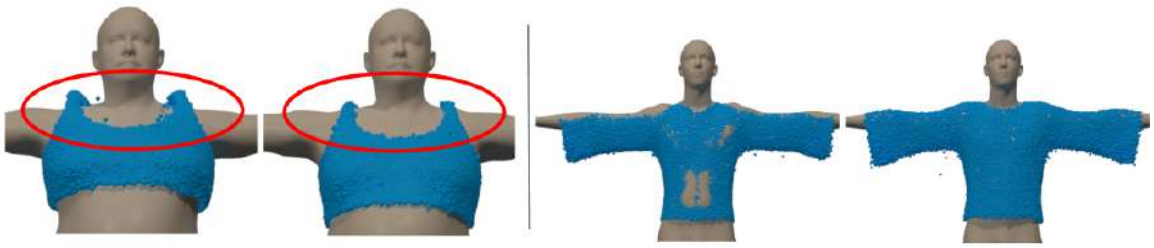


Fig. 6.20 **Impact of body-aware encoding:** Body-aware encoding results in better alignment of clothing and body and also yields less intersection.

From the results presented in Table 6.6 and visualizations shown in Figure 6.20, we observe that our novel body-aware encoding plays a crucial role in improving the alignment between the predicted clothing geometry and the underlying human body. Specifically, this encoding significantly reduces the number of intersections observed between the body mesh and the predicted clothing surface, leading to more physically plausible and visually coherent results. To further examine whether additional constraints could enhance this alignment, we also experiment with training a variant of our model that incorporates the intersection loss proposed in [KKSJ24]. However, we find that this modification does not lead to noticeable improvements in performance, suggesting that our body-aware encoding is already effective in handling such artifacts. Moreover, through detailed qualitative inspection, we find that the misalignment issues—particularly in the form of body-cloth intersections—are more pronounced in the case of upper-body garments, especially in regions around the shoulders. This observation highlights the importance of incorporating body-aware features, as the geometry in these regions tends to be more complex and susceptible to subtle errors in alignment, as shown in Figure 6.20.

6.5 Conclusion

In summary, present TailorDiff, a hierarchical latent point diffusion model for unified and expressive 3D clothing generation. To the best of our knowledge, no prior publicly available work offers a unified model of clothing that is compatible with human models. By leveraging a hierarchical latent space that combines a global shape latent capturing coarse style and a local body-aware latent capturing fine-grained details, our model can generate realistic and detailed clothing shapes that align well with the human body. To achieve this, we introduce a

novel body-aware encoding that enables accurate alignment of cloth and body. TailorDiff demonstrates significant diversity compared to prior work, attributed to both the model and the large-scale dataset. Our approach also enables fine-grained control in generation, which is highly useful for fashion design, editing virtual characters, *etc.* We discuss limitations and a discussion of the proposed method in Section 6.5.1 and future direction in Section 6.5.2.

6.5.1 Limitation and Discussions

Choosing a *learning–friendly* representation which is also agnostic to clothing topology and usable for inverse problems has long been the central focus in clothing research. Recent progress, including our own TailorDiff and contemporaneous work such as LION [ZVW*22] and POCO-LOCO [SDL*25], demonstrates that *point clouds* satisfy these requirements: they side–step mesh parameterisation, scale linearly with resolution, and the explicit representation makes physical modeling easier. Moreover, by treating each point as a point of a 3D Gaussian Splat, a learned cloud can be rendered photorealistically via 3D Gaussian splatting [KKLD23], closing the appearance gap to raster graphics with virtually no architectural change.

This very flexibility comes at a price. Without connectivity, local smoothness is neither enforced nor even well-defined, so noisy point cloud generation has been observed in many prior and concurrent works [ZDW21, ZVW*22]. The classical solution to noisy surface generation involves surface-based Laplacians or normal-consistency losses. However, these cannot be applied directly to an unordered set of point clouds. While post-hoc Poisson reconstruction can regularize the surface (as shown in the results of Section 6.4.5), it breaks the end-to-end learning loop and can result in artifacts ranging from oversmooth features to noisy and peaky surfaces. Designing priors that reconcile the **topological freedom** of points with the **surface regularity** remains an open research question.

One of the challenges in modeling clothing has been selecting a representation that can represent clothing of all topologies and is learning-compatible. Point cloud-based representations are useful, as they can be used to learn a unified model, as shown in this work and other works [ZVW*22, SDL*25]. The very explicitness of points makes them attractive for interaction reasoning, such as collision and contact. Point-cloud-based representation can be used to drive 3D Gaussian Splats [KKLD23] for photo-realistic results.

6.5.2 Future Work

TailorDiff currently conditions on body *shape* only. Extending the latent hierarchy to include *pose* would allow the model to generate garments that drape and wrinkle naturally as the body moves, but doing so demands a large corpus of temporally consistent clothed scans. Curating or synthesising such a dataset—perhaps via cloth simulation engines or video fitting—remains an exciting open problem.

As highlighted in the previous section, point cloud often tends to be noisy. Future work could mitigate these artifacts by (i) fusing distance-field regularisation into the denoising steps, (ii) distilling points into signed-distance or implicit neural surfaces, or (iii) learning a joint point–mesh decoder that captures the surface property of meshes.

TailorDiff represents a first step toward that vision, providing a unified, expressive, and controllable representation of 3-D clothing. As a result TailorDiff marks a first step toward an SMPL-like “foundation clothing model driven by SMPL”, one that supports inverse problems, and human–clothing understanding at scale. Addressing the outlined challenges will be key to turning this vision into a ubiquitous tool for both computer vision and digital human avatars.

II
Human Pose Priors

Human pose space is high-dimensional and heavily constrained by anatomy and inter-joint couplings, so almost every 3D reconstruction pipeline relies on a pose prior to rule out implausible solutions. Data-driven pose priors are trained on large-scale datasets and usually follow Gaussian assumptions in pose input space [BKL*16] or latent space [PCG*19]. In this part of the thesis, we propose a novel formulation for learning a pose prior based on high-dimensional neural distance fields.

In **Chapter 7**, we introduce PoseNDF [TAL*22], which re-frames the problem: instead of learning a probability density, it learns the distance to the manifold of valid poses. The neural distance field gives (i) a precise notion of "how far" an arbitrary pose is from being plausible, (ii) analytic gradients that project noisy or partial estimates back onto the manifold, and (iii) a continuous representation that preserves pairwise pose distances. Empirically, this implicit-manifold prior improves denoising, occlusion completion, and single-image 3D reconstruction, while enabling simple random sampling by drawing quaternions and projecting them to the zero-level set. This also opens more exciting directions for learning the data manifold. The recent follow-up by Ross et al. [RLGCC23] uses neural implicit-based manifold learning beyond articulated bodies and shows results for images.

7

PoseNDF: Modeling Human Pose Manifolds with NDFs

In this chapter, we introduce Pose-NDF [TAL*22], a continuous model for plausible human poses based on Neural Distance Fields (NDFs). Pose and motion priors are critical for generating realistic new poses and accurately reconstructing poses from noisy or partial observations.

Previous research typically learns a lower-dimensional embedding of human poses using Variational Autoencoders (VAEs) and utilizes the learned Gaussian distribution as a prior for reconstruction and pose generation tasks. However, this Gaussian assumption introduces limitations such as bias towards generating more likely poses and sampling from sparsely populated "dead regions."

Pose-NDF models the manifold of plausible human poses directly in the pose space and explicitly as the zero-level set of a neural implicit function, generalizing the concept of implicit surface modeling from 3D geometry to the high-dimensional space of $SO(3)^K$. Here, each human pose is represented by K quaternions, with each quaternion defining rotations on 3-dimensional hyperspheres. The resulting high-dimensional implicit function is fully differentiable with respect to input poses, allowing arbitrary poses to be projected onto the learned manifold through gradient descent optimization directly on hyperspheres. Unlike VAE-based pose priors that warp pose space into a Gaussian latent distribution, our approach maintains the actual manifold structure of the pose space, preserving intrinsic distances between poses. As a result, Pose-NDF is free from the restrictive Gaussian assumptions and avoids related limitations such as mode collapse or generation biases.

We demonstrate that Pose-NDF consistently outperforms existing state-of-the-art methods across various downstream tasks, including denoising real-world motion capture data, pose recovery from occluded observations, and accurate 3D pose reconstruction from images. Furthermore, our method excels at generating a more diverse set of plausible poses through random sampling and subsequent projection onto the learned manifold compared to traditional VAE-based approaches.

This chapter is based on the work titled **Pose-NDF: Modeling Human Pose Manifolds with Neural Distance Fields, Tiwari *et al.*, ECCV 2022** [TAL*22]. The associated code, models, and datasets have been publicly released for research. This work received a *Best Paper Honourable Mention* award at ECCV 2022 and has been cited over 100 times since publication. Subsequent research efforts have expanded upon this foundational idea, applying Neural Distance Fields to manifold learning with Riemannian distances [HTB*24],

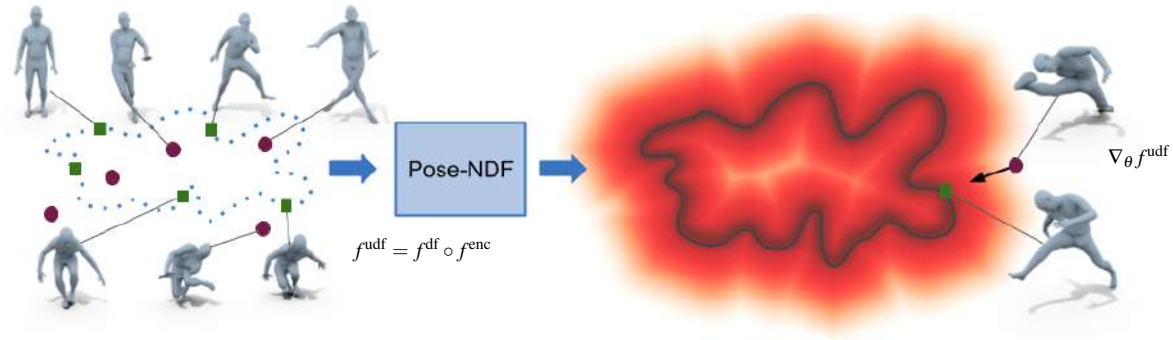


Fig. 7.1 **Pose-NDF-a neural unsigned distance field in $SO(3)^K$ space:** Pose-NDFlearns the manifold of plausible poses as zero level set. We learn the distance field representation from samples of plausible (■) and unrealistic (●) poses (*left*). We encode the input pose (given as a set of quaternions) using a structural MLP f^{enc} and predict the distance from the joint representation using an MLP f^{df}. The gradient $\nabla_\theta f^{\text{udf}}$ and distance value $f^{\text{udf}}(\theta)$ are used to project implausible poses onto the manifold (*right*).

developing robust motion models [DFZ*24], and exploring 2D-pose priors based on polar coordinates [GNF*25].

7.1 Introduction

Realistic and accurate human motion capture and generation are essential for understanding human behavior and human interaction in the scene [GMSPM21, vMHB*18b, ZHN*20, ZZM*20, BXP*22]. Human motion capturing systems, including optical marker-based systems [LMB14b, KMPA21], IMU-based methods [GMSPM21, vMHB*18b], or methods based on reconstruction from RGB/RGB-D data [SLNW16, GZW*20, JXY*18, ZZQ*20], frequently encounter problems such as skating artifacts, self-intersections, jitters, and generally produce unrealistic human poses, particularly in the presence of noisy data or occlusion. These limitations hinder the direct applicability of captured motion data in critical applications such as 3D scene understanding, human motion generation, and AR/VR applications. To overcome these problems, researchers, designers, and engineers often resort to exhaustive manual or automatic post-processing procedures, adding significant complexity and time overhead to motion data handling.

In recent years, learned data priors for post-processing non-realistic human poses have gained increasing popularity. Prior human pose models mainly focus on learning joint distributions of individual body joints directly in pose space [BKL*16, ENNE12, AB15, BB00] or, more recently, in a latent space using Variational Autoencoders (VAEs) [PCG*19, RBH*21, ZZB*21]. While these methods have significantly enhanced plausibility and realism of poses after model fitting, VAE-based approaches such as VPoser [PCG*19] and HuMoR [RBH*21] inherently rely on a Gaussian assumption within the latent pose space. This assumption introduces critical limitations:

- They inherently bias pose generation towards poses close to the mean of the Gaussian, potentially disregarding correct yet less frequently observed poses.
- The latent spaces in VAEs do not preserve distances between individual poses, causing small movements in latent space to correspond disproportionately to large and often unrealistic movements in pose space.
- VAEs tend to "fold" the manifold into a Gaussian distribution [MSJG16], creating "dead regions" in the outer areas of the distribution devoid of realistic poses, leading to the generation of implausible samples far from the input pose, as shown in our experimental analysis.

To alleviate these issues, we present Pose-NDF, a novel human pose prior explicitly designed to represent the complete manifold of plausible poses within a high-dimensional pose space. Our key insight is to model this manifold as a surface where plausible human poses exactly lie on the surface of the manifold, represented by zero distance, and non-plausible poses reside outside, indicated by non-zero distances. To learn this manifold, we propose using a high-dimensional neural field, analogous to neural distance fields that have been successfully applied to represent complex 3D shapes [PFS*19b, CMPM20]. This novel representation maintains explicit distances between poses, enabling traversal in pose space guided by the negative gradient of the distance function, inherently pointing toward the direction of maximum distance decrease. By performing gradient descent starting from potentially unrealistic and non-plausible poses, our approach robustly locates the nearest plausible pose on the manifold.

An overview of our method is given in Figure 7.1. Specifically, we formulate the problem of learning a pose manifold as a surface completion task in n-dimensional space. In order to learn a pose manifold in this high-dimensional space, there are two key challenges:

- The inherent high dimensionality of the input pose space, *i.e.* $K \times d$, where $K = 21$ and $d = 4$ in our case. In contrast to 3-dimensional space for learning 3D shapes, each point in our proposed manifold belongs to $\mathbb{R}^{K \times d}$.
- The non-Euclidean nature of the input space, unlike traditional 3-dimensional implicit surfaces [PFS*19b, CMPM20]. Instead, the pose space is given as $SO(3)^K$, in which a single pose can be represented by K elements of the rotation group $SO(3)$, describing the orientations of joints in a human body model.

To represent group elements, we opted for a quaternion representation, as they are continuous, have an easy-to-compute distance, and are subject to an efficient gradient descent algorithm. We apply a hierarchical implicit neural function that maps any given pose to its corresponding distance based on the kinematic structure of the human body [MZBT21]. We train our model using the AMASS dataset [MGT*19a], where each sample from the dataset is treated as a point on the manifold.

The learned neural field representation can be used to generate poses or denoise poses by projecting arbitrary poses onto the plausible pose manifold, similarly to methods described in [CMPM20]. Leveraging this property, we extensively demonstrate Pose-NDF's effectiveness across diverse applications, including pose interpolation, diverse and realistic pose

generation, pose priors for accurate 3D pose estimation from images [PCG*19, BKL*16], and robust motion denoising [GMSPM21, RBH*21]. Our experiments show that our proposed approach consistently outperforms existing state-of-the-art methods across all these applications.

In summary, our contributions are:

- A novel high-dimensional neural distance field representation in $SO(3)^K$, Pose-NDF, which represents the manifold of plausible human poses. Pose-NDF can be easily used to clean any arbitrary pose by projecting onto the learned manifold, using the distance field and its gradient.
- Pose-NDF improves the state of the art in human body fitting from images by acting as a pose prior. It outperforms other human pose priors, such as VPoser [PCG*19] and the human motion prior HuMoR [RBH*21] on motion denoising.
- Our method is comparable or faster than current state-of-the-art methods, is fully differentiable, and the distance from the manifold can be leveraged for finding the optimal step size during optimization.
- Pose-NDF generates more diverse and realistic samples compared to traditional Gaussian-based methods, which often bias towards generating commonly observed or average poses. Our method, by explicitly modeling the entire plausible manifold, promotes diversity and realism in generated poses, thereby enhancing the practical applicability of our approach across multiple scenarios.

We begin by introducing our approach in Section 7.2, where we describe the pose representation used in our model. We introduce Hierarchical Implicit Neural Function, which we use to learn our manifold. This network leverages a human body kinematics-aware encoder, built around the human skeletal hierarchy, and an MLP decoder that learns to regress a distance-based field from the given pose features. This design enables our model to effectively capture the tree-based hierarchical structure of the human skeleton.

In Section 7.3, we discuss a more general perspective, demonstrating how neural distance field formulations can be used to model the data manifold of any generic data, such as images. This formulation is not only theoretically elegant but also opens up promising new directions, including conceptual links to denoising diffusion models and energy-based generative modeling.

We present experimental results in Section 7.4, highlighting the advantages of our model, Pose-NDF, as an efficient pose prior across various downstream tasks. These include motion denoising, monocular 3D human pose estimation from images, and generative sampling of plausible human poses. We compare our method against several state-of-the-art baselines and show improvements. Finally, in Section 7.5, we discuss the current limitations of our method and provide a brief discussion around future work, which solves these limitations.

7.2 Method

In this section, we describe our method Pose-NDF, a model for manifolds of plausible human poses based on high-dimensional neural fields. We assume that the realistic and plausible human poses lie on a manifold embedded in pose space $SO(3)^K$, with K being the number of joints in the human body. Given a neural network $f : SO(3)^K \mapsto \mathbb{R}^+$, which maps a pose, $\theta \in SO(3)^K$ to a non-negative scalar, we represent the manifold of plausible poses as the zero level set:

$$\mathcal{S} = \{\theta \in SO(3)^K \mid f(\theta) = 0\}, \tag{7.1}$$

such that the value of f represents the unsigned distance to the manifold, similar to neural fields-based 3D shape learning [CLI*20, CMPM20, GYH*20, PFS*19b]. Without loss of generality, we use the SMPL body model [LMR*15, PCG*19], resulting in poses θ with $K = 21$ joints.

7.2.1 Quaternions as Representation of SO(3)

A human pose is represented by 3D rotations of individual joints in the human skeleton. The 3-dimensional rotation group $SO(3)$ has several common vector space representations that are used to describe group elements in practice. Frequently used examples are rotation matrices, axis-angle representations, or unit quaternions [Huy09]. Pose-NDF requires the representation to have specific properties:

- We aim to model a *manifold*, continuously embedded in pose space. Thus, the chosen representation should be continuous in parameter space, which prohibits axis-angle representations
- The representation should enable efficient computation of the geodesic distance between two elements
- Our algorithm requires *efficient gradient descent* in pose space.

As described in Section 7.2.4, quaternions are subject to such a gradient descent algorithm that makes use of the efficient reprojection to $SO(3)$ by vector normalization. In contrast, rotation matrices would require more expensive orthogonalization. Therefore, we chose unit quaternions as the best-suited $SO(3)$ representation of joints, as they fulfill all three properties. We will use $\theta = \{\theta_1, ..., \theta_K\}$ to denote the quaternions for all K joints of a pose. Each quaternion represents the rotation of a joint with respect to its parent node. Since quaternions lie on S^3 (embedded in 4-dimensional space) the full pose θ can be easily used as input for a neural network $f : \mathbb{R}^{4K} \to \mathbb{R}^+$. We define the distance $d : (S^3)^K \times (S^3)^K \to \mathbb{R}^+$ between two poses $\theta = \{\theta_1, ..., \theta_K\}$ and $\hat{\theta} = \{\hat{\theta}_1, ..., \hat{\theta}_K\}$ as:

$$d(\theta, \hat{\theta}) = \sqrt{\sum_{i=1}^{K} \frac{w_i}{2} (\arccos |\theta_i^\top \cdot \hat{\theta}_i|)^2}, \tag{7.2}$$

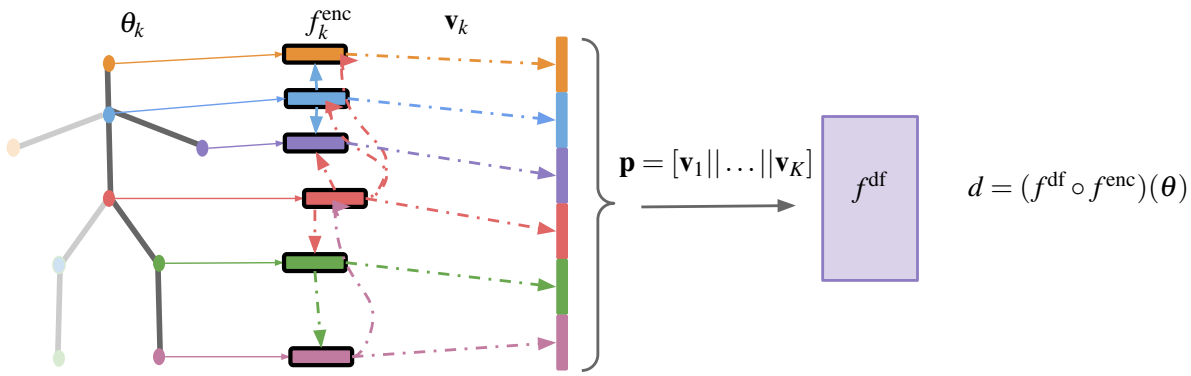


Fig. 7.2 **Hierarchical Implicit Neural Function:** $f^{\text{udf}}(\theta) = (f^{\text{df}} \circ f^{\text{enc}})(\theta)$ First we encode the pose θ_k using f_k^{enc}, where $k \in \{2 \ldots K\}$. Each joint pose is embedded using an MLP f_k^{enc}, conditioned on the parent features $\mathbf{v}_{\tau(k)}$. The resulting embeddings $\mathbf{v}_k$ are concatenated to form the pose representation $\mathbf{p}$, which is processed by f^{df} to predict unsigned distance fields. For visuslisation, we use a skeleton with less joints.

where the individual elements of summation are a metric on $SO(3)$ [Huy09] and w_i is the weight associated with each joint based on their position in the kinematic structure of the SMPL body model (*i.e.* early joints in the chain have higher weights). It should be noted that the double cover property of unit quaternions, that is, the quaternions $\mathbf{q}$ and $-\mathbf{q}$ represent the same $SO(3)$ element, does not lead to additional challenges. We simply train the network to be point symmetric by applying sign flip augmentation on input quaternions.

7.2.2 Hierarchical Implicit Neural Function

We represent the human pose using quaternions in the local coordinate frames of the parent joint, leveraging the kinematic structure of the SMPL body model. We treat the joints in a local coordinate frame, so that continuous manipulation of a single joint corresponds to realistic motion. However, this might result in an unrealistic combination of joint rotation. As discussed in prior works [AKH19], incorporating the kinematic structure of the human skeleton improves the performance for pose modelling tasks. The plausibility of individual joints depends on the rotations of their parent joints and thus needs to be conditioned on them. In order to incorporate this dependency, we use a hierarchical network f^{enc}, which encodes the human pose based on the model structure [AKH19, GLK*20, MZBT21], before predicting the distance based on the joint representation.

Formally, for a given pose $\theta = \{\theta_1, \ldots, \theta_K\}$, where θ_k is the pose for joint k, and a function $\tau(k)$, mapping the index of each joint to its parent joints index, we encode each pose using an MLP as:

$$f_1^{\text{enc}} : (\theta_1) \mapsto \mathbf{v}_1 \quad f_k^{\text{enc}} : (\theta_k, \mathbf{v}_{\tau(k)}) \mapsto \mathbf{v}_k, \quad k \in \{2 \ldots K\} \tag{7.3}$$

which takes the quaternion pose and encoded feature $\mathbf{v}_{\tau(k)} \in \mathbb{R}^l$ of its parent joint as input and generates $\mathbf{v}_k \in \mathbb{R}^l$, where l is the dimension of feature. We then concatenate the encoded feature for every joint to get a combined pose embedding $\mathbf{p} = [\mathbf{v}_1 || \ldots || \mathbf{v}_K]$. This embedding is processed by an MLP $f^{\text{df}} : \mathbb{R}^{l \cdot K} \rightarrow \mathbb{R}^+$, which predicts the unsigned distance for the given pose representation $\mathbf{p}$. Collectively the complete model $f^{\text{udf}}(\theta) = (f^{\text{df}} \circ f^{\text{enc}})(\theta)$, is termed as Pose-NDF, where $f^{\text{udf}} : SO(3)^K \mapsto \mathbb{R}^+$. We show the detailed network architecture in Figure 7.2.

7.2.3 Loss functions

We train the hierarchically structured neural field f^{udf} to predict the geodesic distance to the plausible pose manifold for a given pose. The training data is given as a set $\mathcal{D} = \{(\theta_i, d_i)\}_{1 \leq i \leq N}$, containing pairs of poses θ_i and distances d_i (Eq. 7.2). We train the network with the standard distance loss $\mathcal{L}_{\text{UDF}}$, and an Eikonal regularizer $\mathcal{L}_{\text{eikonal}}$, which encourages a unit-norm gradient for the distance field outside of the manifold [CL83, GYH*20]:

$$\mathcal{L}_{\text{UDF}} = \sum_{(\theta,d)\in\mathcal{D}} ||f^{\text{udf}}(\theta) - d_\theta||_2 \tag{7.4}$$

$$\mathcal{L}_{\text{eikonal}} = \sum_{(\theta,d)\in\mathcal{D},\, d\neq 0} (||\nabla_\theta f^{\text{udf}}(\theta)|| - 1)^2, \tag{7.5}$$

7.2.4 Sampling via Projection on Manifold

Given a trained model f^{udf}, it can be applied to project an arbitrary pose θ to the manifold of plausible poses. We use the predicted distance $f^{\text{udf}}(\theta)$ and gradient information $\nabla_\theta f^{\text{udf}}(\theta)$ to project a query pose to the manifold surface $\mathcal{S}$, as was previously done in unsigned distance functions for 3D shapes [CMPM20]. In our case, given $SO(3)$ poses, this amounts to finding:

$$\hat{\theta} = \underset{\theta\in SO(3)^K}{\arg\min}\, d(\theta, \mathcal{S}), \tag{7.6}$$

where $d(\theta, \mathcal{S})$ is the distance (Eq. 7.2) of θ to the closest point in $\mathcal{S}$. We find $\hat{\theta}$ by applying gradient descent on the 3-sphere, using gradient information $\nabla_\theta f(\theta)$ and distances $f(\theta)$, obtained from the implicit neural function f. One step is given as:

$$\theta^i = \theta^{i-1} - \alpha f(\theta^{i-1}) \nabla_\theta f(\theta^{i-1}), \tag{7.7}$$

followed by a re-projection to the sphere (*i.e.*, vector normalization) after several iterations. This algorithm is guaranteed to converge to local minima on the sphere, which, in our case, assuming a correctly learned distance function, is the nearest point on the pose manifold.

7.2.5 Data Preparation

In this section, we explain the data preparation for training and evaluation of our model.

Training data: We use the AMASS dataset [MGT*19a] to train our model. AMASS (Archive of Motion Capture as Surface Shapes) dataset unifies multiple mocap datasets [Car, HTRM15, Adv, GMT*20, SBB10] into a common format of realistic human body meshes using SMPL. It provides diverse motion sequences with consistent topology, ideal for learning human pose and shape models. We assume that AMASS represents a dataset of realistic poses, and all the poses that are not part of AMASS are not plausible. As mentioned in the experiment section, we prepare a dataset of points (poses) and corresponding distance values using kNN. We implement kNN using FAISS [JDJ19] and Pytorch3D [RRN*20]. We first approximate k'-NN of a query pose using FAISS and L2 distances, where $k' >> k$. Then we use the geodesic distance to find exact k neighbours from these k' neighbours. Specifically, we use $k' = 500$ and $k = 5$ in our case. We use this multi-step approach because the AMASS dataset is very large and our proposed data preparation steps are efficient. We generate the ground truth distance by taking the average of the k smallest distances. We use approximately 21M poses from the AMASS dataset for data preparation and training. Following VPoser [PCG*19], we also sample random frames from the motion sequence data, in order to avoid repetitive poses.

Evaluation and Validation: For validation, we use the validation split of the AMASS dataset and prepare distance values with respect to the train split. For testing the accuracy of the distance field prediction in our model, we use the test split of AMASS dataset. For downstream tasks, we use existing real-world mocap data like [GMSPM21] and [MGT*19a]. For comparison on the EHF dataset, we follow the evaluation provided in [PCG*19], *i.e.* we align the predicted SMPL mesh with the ground truth using Procrustes and then calculate the error on SMPL vertices (not on face and hands).

7.2.6 Training and Implementation details

We illustrate the detailed architecture of our network in Figure 7.2. The hierarchical pose encoding module is implemented using structural MLPs, inspired by prior work such as LEAP [MZBT21] and Aksan et al. [AKH19]. Each structural MLP comprises two fully connected layers, designed to capture local pose-dependent features effectively. These features are then passed to a subsequent 5-layer MLP, which is responsible for predicting the implicit distance field representation. For all hidden layers [GYH*20], we use the softplus activation function with a sharpness parameter $\beta = 100$, which ensures smooth gradients while maintaining the capacity to model high-frequency details.

7.3 Learning and Sampling Data Manifold

In PoseNDF [TAL*22], we present a novel formulation based on neural distance fields to learn a data manifold that can be used for generation and as a prior. Recent work, published after PoseNDF, has extended the formulation by applying Neural Distance Fields to manifold learning with Riemannian distances [HTB*24], developing robust motion models [DFZ*24], and exploring 2D-pose priors based on polar coordinates [GNF*25]. However, they are still focused on human poses, motion, and articulation. In this section, we discuss some contemporary works that move beyond the task of human articulation and closely resemble the neural distance field-based formulation for manifold learning or generative models. We also provide a brief discussion around similarities between diffusion model sampling and PoseNDF sampling in Section 7.3.2.

7.3.1 Energy Based Implicit Manifold

Most generative models sample $z \sim p_{\text{base}}$ in $\mathbb{R}^m$ and push it through a network f_θ to get $x = f_\theta(z)$. Because any continuous $f_\theta : \mathbb{R}^m \to \mathbb{R}^n$ is a single chart, such push-forward models cannot represent data that live on manifolds with holes, self–intersections or multiple components without tearing or exploding Jacobians, which harms likelihoods and numerical stability.

EBIM [RLGCC23] eliminates this topological bottleneck by learning the manifold as a zero-set instead of a deformation of Euclidean space and then learning density within this neural implicit manifold by fitting a constraint Energy-based model. The idea behind representing the data manifold as the zero set of a neural network is similar to the idea presented in PoseNDF. In PoseNDF, we assume that the learned manifold lies in the input space and hence $M \subset \mathbb{R}^n$ $(m = n - 1)$, whereas in EBIM, the submanifold $M \subset \mathbb{R}^n$ $(m < n)$. Moreover, PoseNDF is trained with both positive and negative samples. But EBIM only needs positive samples. Moreover, EBIM also provides density estimates on the learned manifold. However, PoseNDF only models the distance field and doesn't capture the probabilty density of data points.

We find it interesting to see that the neural implicit manifold-based formulation has been extended to learning generative modeling for generic datasets, such as images, in [RLGCC23].

7.3.2 Sampling in NDF and Diffusion

In this section, we examine the similarities between the sampling mechanisms employed by two contemporary generative paradigms. Although both methods learn a scalar field whose gradient is normal to the data manifold, they place fundamentally different objectives on that field and thus arrive at distinct, yet related sampling routines. The two techniques occupy complementary ends of the spectrum: diffusion treats sampling as stochastic simulation of a learned SDE, whereas PoseNDF treats it as deterministic geometric projection.

Here, we specifically discuss the denoising process used in the LDM [RBL*21], which generates images through an **iterative denoising process**. The model starts with a random

Gaussian noise latent and repeatedly refines it into a coherent image latent via a learned denoising U-Net. At each step of the reverse diffusion, the U-Net **predicts the noise component** in the current noisy latent, which is then subtracted from the latent. By gradually removing noise according to a scheduled sequence of noise levels (from high to low), the latent becomes increasingly structured, ultimately yielding a clean final latent that a decoder maps to an output image.

This sequential sampling procedure, based on a Markov chain of denoising steps, is the crux of Stable Diffusion's or any similar diffusion model's sampling strategy. Each denoising step moves the sample closer to the data manifold of real images and can be seen as a learned **projection of a noisy sample onto the image distribution manifold**. This is similar to how we formulate sampling via projection in PoseNDF. Formally, if $z_T \sim \mathcal{N}(0, I)$ is the initial random latent at the T-th (noisiest) diffusion step, the model applies an update of the form:

$$z_{t-1} = z_t - f_\theta(z_t, t),$$

where $f_\theta(z_t, t)$ is a function of the U-Net's noise prediction (scaled by the diffusion noise schedule) at step t. Over T iterations, these updates yield z_0, a noise-free latent. In practice, techniques such as DDIM or Euler samplers implement this recurrence using specific formulas for f_θ and optional stochasticity. Instead of a diffusion process, PoseNDF represents the set of valid data as the zero-level set of a neural implicit function (a neural distance field) in the high-dimensional parameter space. The projection algorithm mentioned in Equation (7.7), is exactly similar to the sampling mentioned above. The projection onto the manifold in PoseNDF is also performed iteratively, and instead of denoising data based on the noise level, we utilize the term distance from the manifold. In essence, each can be seen as progressively "denoising" or correcting the sample: Stable Diffusion explicitly removes noise to reveal an image, while PoseNDF removes implausibility (distance to manifold) to yield a valid pose.

7.4 Experiments and Results

In this section we evaluate Pose-NDF and show the different use cases of our pose model, which include the ability to serve as a *prior in denoising motion sequences or recovery* from partial observations (Section 7.4.2), *prior for recovering plausible poses* from images (Section 7.4.4) using an optimization-based method, *pose generation* (Section 7.4.5) and *pose interpolation* (Section 7.4.6). We demonstrate that the Pose-NDF method *outperforms the state-of-the-art VAE-based human pose prior methods*. We also show the *advantages* of our distance field formulation over VAEs or Gaussian assumption models (Section 7.4.7). Before presenting the results, we provide an explanation of the training and implementation details of Pose-NDF in Section 7.4.1.

Data	HPS [GMSPM21]			AMASS [MGT*19a]			Noisy AMASS		
# frames	60	120	240	60	120	240	60	120	240
Method									
VPoser [PCG*19]	4.91	4.16	3.81	1.52	1.55	1.47	8.96	9.13	9.15
HuMoR [RBH*21]	9.69	8.73	10.86	3.21	3.62	3.67	11.04	17.14	30.31
Pose-NDF	**2.32**	**2.14**	**2.11**	**0.59**	**0.55**	**0.54**	**7.96**	**8.31**	**8.46**

Table 7.1 **Motion denoising**: We compare the per-vertex error (in *cm*) on mocap data from HPS (*left*) and AMASS (*middle*) and on artificially created noisy AMASS data (*right*). In all cases, Pose-NDF based motion denoising results in the least error. We also observe that in the case of mocap data (HPS, AMASS), motion denoising using Pose-NDF results in a very small error (a small change from the input), which is the desired behavior, as these mocap poses are already realistic and hence close to our learned manifold. On the other hand, HuMoR significantly changes the input pose.

7.4.1 Experimental Setup

We use the AMASS dataset [MGT*19a] for training. As mentioned in Section 7.2.3, we train the network with supervision on predicted distance values, thereby creating a dataset of pose and distance pairs (θ, d_θ). Since the training samples from AMASS lie on the desired manifold, $d_\theta = 0$ is assigned to all poses in the dataset. We then randomly generate negative samples with distance $d_\theta > 0$ by adding noise to AMASS poses. We train our model in a **multi-stage** regime by varying the type of training samples used. We start our training with manifold poses θ_m and non-manifold poses θ_{nm} with a large distance to the desired manifold. Then we increase the number of non-manifold poses θ_{nm} with a small distance in each training batch. This training scheme helps to initially learn a smooth surface and to iteratively introduce the fine details over the course of training. Our **network architecture** consists of one 2-layer MLP f^{enc} with an output feature size of $l = 6$ for each joint, similar to [MZBT21]. Thus, the pose encoding network generates a feature vector $\mathbf{p} \in \mathbb{R}^{126}$. We implement the distance field network f^{df} as a 5-layer MLP. For training, we use the softplus activation in the hidden layer and train the network end-to-end using the loss functions described in Equation (7.4).

7.4.2 Denoising Mocap Data

Human motion capture has been performed using diverse setups, ranging from RGB to RGB-D and IMU-based capture systems. The data captured from these sources often produces artifacts, such as jitters, unnaturally rigid joints, or unusual bends at certain joints or positions with only partial observations. Prior work [RBH*21] enhances the quality of captured motion sequences by using an optimization-based method, aiming to recover the captured data while preserving the realism of human poses. A robust and expressive human pose prior is crucial for preserving the realism of optimized poses, while also maintaining the original

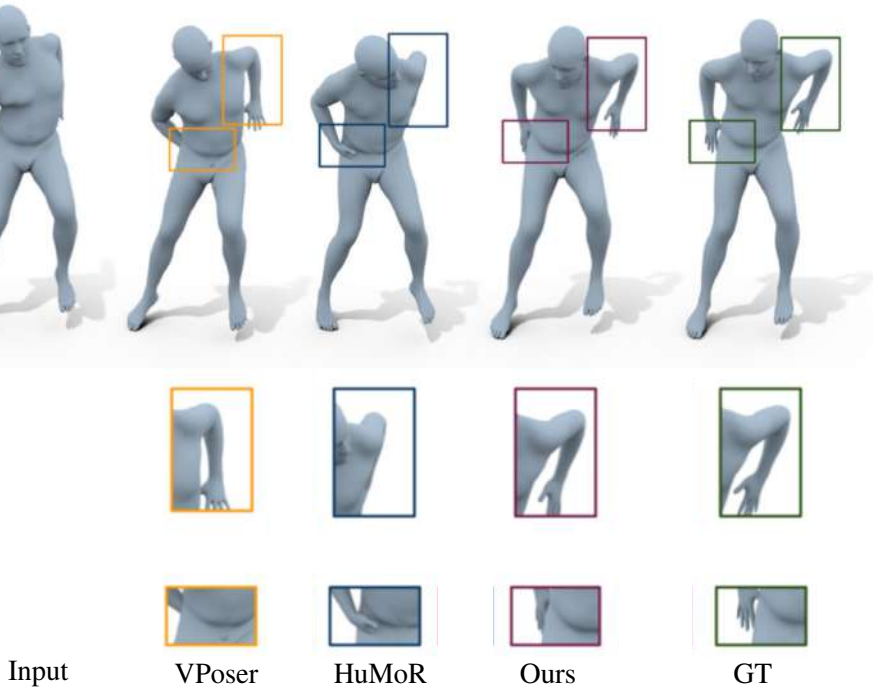


Fig. 7.3 **Comparison with prior works on motion denoising**: We observe that Pose-NDF based motion denoising makes the pose realistic and solves small intersection issues, while VPoser and HuMoR still result in unrealistic poses.

data. Following HuMoR [RBH*21], we demonstrate the effectiveness of our pose manifold for motion denoising in this section and fitting to partial observation in the next section.

We follow the same experimental setup as [RBH*21], but only deal with human poses and thus, remove the terms corresponding to human-scene contact and translation of root joint. In total, we find the pose parameters $\hat{\theta}^t$ at frame t as:

$$\hat{\theta}^t = \arg\min_{\theta} \lambda_{\mathrm{v}} \mathcal{L}_{\mathrm{v}} + \lambda_{\theta} \mathcal{L}_{\theta} + \lambda_t \mathcal{L}_t \tag{7.8}$$

where $\mathcal{L}_{\mathrm{v}}$ makes sure that the optimized pose is close to the observation and the temporal smoothness term $\mathcal{L}_t$ enforces temporal consistency:

$$\mathcal{L}_{\mathrm{v}} = ||\mathcal{J}(\beta_0, \hat{\theta}^t) - \mathcal{J}_{\mathrm{obs}}||_2^2 \tag{7.9}$$

$$\mathcal{L}_t = ||M(\beta_0, \hat{\theta}^{\mathrm{t}}) - M(\beta_0, \theta^{\mathrm{t-1}})||_2^2 \tag{7.10}$$

Here, $\mathcal{J}(\beta, \theta)$ represent vertices (mocap markers) and $M(\beta, \theta)$ represents SMPL mesh vertices for a given pose (θ) and shape (β) parameters of SMPL [LMR*15]. Finally, we use Pose-NDF as a pose prior term in the optimization by minimizing the distance of the current pose from our learned manifold, $\mathcal{L}_{\theta} = f^{\mathrm{udf}}(\theta)$. We leverage the distance $f^{\mathrm{udf}}(\theta)$ to get the optimal step size during optimization.

Experimental setup for denoising using VPoser: VPoser does not have motion experiments, which is why we combine the latent space optimization from the original work with our optimization given in Equation (7.8) to ensure that we compare against the best possible result. Specifically, we first encode the rotation matrix representation of noisy input pose θ^t using the VPoser encoder as $\mathbf{z}^t = f_{\text{v_enc}}(\theta^t)$, then add random noise ($\hat{\varepsilon}$) in the latent space and reconstruct the pose by $\tilde{\theta}^{\text{t}} = f_{\text{v_dec}}(\mathbf{z}^t + \hat{\varepsilon})$. Following [ZZB*21], we observe that the temporal term in latent space yields better results than the temporal term in input pose/vertices, which we used in the VPoser experiment. The prior and temporal term for VPoser-based denoising are given as:

$$\mathcal{L}_{\theta}^{\text{VPoser}} = ||\hat{\varepsilon}||_2 \tag{7.11}$$

$$\mathcal{L}_{t}^{\text{VPoser}} = ||\mathbf{z}^{t-1} - \hat{\mathbf{z}}^{t}||_2. \tag{7.12}$$

We evaluate in two different settings:

- **Refining mocap dataset:** Real-world mocap dataset might suffer from jitters, unnatural mapping between IMU skeleton and SMPL skeleton. With this setup, we aim to refine the mocap data to look more plausible and realistic.
- **Denoising noise-injected mocap dataset:** For fair comparison and understanding the bounds of our model, we follow the setup used in [RBH*21], where we add noise to the mocap dataset and use our pose prior to denoise the motion.

Refining mocap dataset

For clean mocap datasets, we use HPS [GMSPM21] and the test split of AMASS [MGT*19a, PCG*19]. We use a list of SMPL mesh vertices $\mathcal{J}$ as observation during optimization. We created the data with a fixed shape and do not optimize for shape parameters β. Instead of adding noise to joint locations, we add noise directly to the rotation of each joint. This is done for all methods to ensure a fair comparison. For HuMoR [RBH*21], we use the *TestOpt* optimization from the original work. For VPoser, we used the setup explained before.

We compare motion denoising between HuMoR [RBH*21] (*TestOpt*), Equation (7.8) with VPoser prior [PCG*19], and Equation (7.8) with Pose-NDF prior in Table 7.1, under the column named "*HPS [GMSPM21]*" and "*AMASS [MGT*19a]*". We observe that Pose-NDF achieves the lowest error in all settings. For mocap datasets like AMASS and HPS, the motion is realistic, but can have small artifacts and jitter. Thus, an ideal motion/pose prior should not change the overall pose of these examples, but only fix these local artifacts. We observe that, numerically, VPoser and Pose-NDF-based optimization do not change the input pose significantly. However, HuMoR changes the pose, and this change increases with an increasing number of frames. This is because HuMoR is a motion-based prior (conditioned on the previous pose) and, hence, over time, the correction in pose accumulates and makes the output pose significantly different from the input.

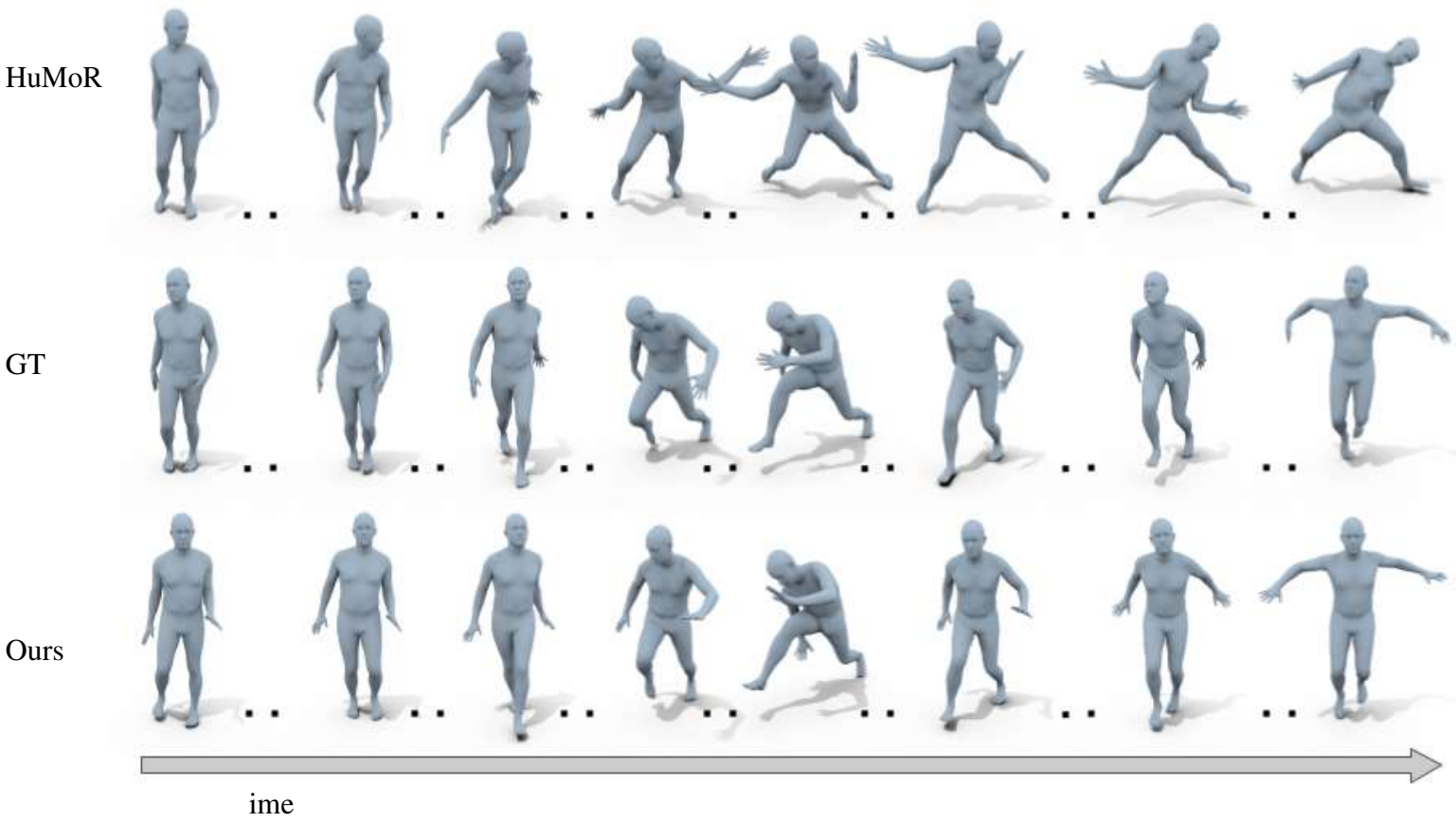


Fig. 7.4 **Motion denoising using Pose-NDF v/s HuMoR [RBH*21]**: We compare the results on motion denoising using Pose-NDF and HuMoR [RBH*21] as priors with GT data and visualize every 10^{th} frame of a sequence. We observe that for HuMoR (*top*), the correction in input pose accumulates over time and makes the output pose significantly different from the GT (*middle*). Pose-NDF remains close to observations while correcting unrealistic poses (*bottom*).

Method \ Data	HPS	AMASS	Noisy AMASS	Partial Observation
Input	3.13	1.14	2.54	-
VPoser [PCG*19]	3.16	1.36	2.54	2.98
HuMoR [RBH*21]	2.81	**1.04**	1.43	2.13
Pose-NDF	**2.28**	1.05	**1.40**	**2.01**

Table 7.2 **Pose realism using self-intersection:** Average % of self-intersecting mesh faces in motion denoising. It can be seen that in most cases, Pose-NDF produces fewer intersections than VPoser and HuMoR.

Denoising noisy motion data

For the noisy mocap dataset, we create random noisy sequences by adding Gaussian noise to AMASS test sequences and call it “Noisy AMASS”. The average noise introduced in “Noisy AMASS” is 9.3 *cm*. For the “Noisy AMASS" data, Pose-NDF-based optimization outperforms prior work. We visualise the denoising results in Figure 7.3, and observe that the Pose-NDF-based method produces realistic and close to GT results. We further compare results of a sequence with HuMoR in Figure 7.4. HuMoR results in large deviations from the input/GT, due to the accumulation of corrections over time. We compare motion denoising between HuMoR [RBH*21] (*TestOpt*), Equation (7.8) with VPoser [PCG*19], and Equation (7.8) with Pose-NDF in Table 7.1, under the column "*Noisy AMASS*".

Similar to the observation made for refining the mocap dataset, Pose-NDF achieves the lowest error in this setting as well. The error accumulation for HuMoR results in lower performance for HuMoR. We show more results of motion denoising in Figure 7.5. We observe that the VPoser-based motion denoising does not change the pose much if the noisy observation seems similar to common poses like standing with hands near the body (first row in Figure 7.5) and changes significantly in a random way when the poses look rare (bottom row in Figure 7.5). HuMoR changes the pose into more realistic poses at first (for initial frames), but results in unrealistic poses later because of the accumulation of changes over time, *e.g.* (bottom-left in Figure 7.5). Pose-NDF seems to perform better in all such cases, resulting in realistic poses and also does not deviate much from noisy observations.

We evaluate the average percentage of self-intersecting mesh faces in the motion denoising task to evaluate which method produces more realistic poses. As seen from qualitative examples, HuMoR and VPoser generate results with self-intersecting poses. We evaluated this quantitatively, by counting the number of intersecting faces for all methods in Table 7.2. Pose-NDF clearly produces fewer intersections, which we credit to our detailed manifold.

7.4.3 Fitting to Partial Data

As mentioned, we can also use our prior in an optimization-based setup for recovering the whole SMPL mesh, given partial observation. We use the test set of AMASS and randomly create occluded poses (*e.g.* missing arm or legs or shoulder joint) and quantitatively

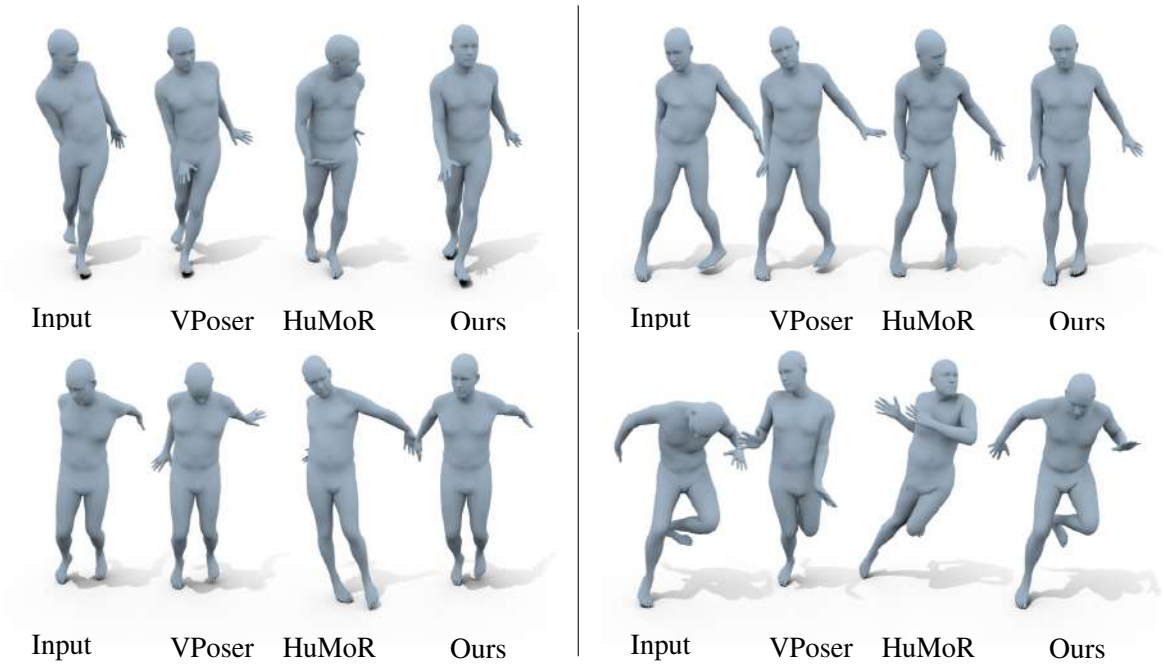


Fig. 7.5 **Motion denoising for noisy mocap data ("Noissy AMASS"):** We observe that Pose-NDF based motion denoising makes the pose realistic and also resembles the input observation, while VPoser and HuMoR still result in unrealistic poses and, in some cases, deviate too much from the input observation (*e.g. bottom-left*).

compare with HuMoR [RBH*21] and VPoser [PCG*19] in Table 7.3. We use Equation (7.8) for VPoser and Pose-NDF-based optimization. However, we modify the optimization objective and variables. We only optimize for the occluded joints, and for our model, we initialize the occluded joint pose randomly (close to 0). Regarding optimization objective, for Equation (7.9) and Equation (7.10), we only use the visible joints and vertices part of visible parts(bones), *e.g.*, for the occluded right hand, we excluded all the vertices which belong to the body parts upper arm, lower arm, and hand. For HuMoR, we use the *TestOpt* provided in their paper. We evaluate three different types of occlusions:

- **Occluded left leg:** To understand the bias of prior towards generating standing poses.
- **Occluded left arm:** To evaluate the diversity of pose generated.
- **Occluded right shoulder and upper arm:** To evaluate the plausibility of the pose generated.

Each of these setups is designed to evaluate a specific aspect. Since the dataset is more likely to contain standing poses, we use this setup to evaluate the bias introduced by the data. For the occluded arm case, it is easier to generate plausible, non-intersecting poses; hence, we primarily assess diversity using this setup. In the occluded shoulder+arm case, we argue that even a small deviation in the shoulder joint can cause an unrealistic bend, leading to more incorrect and intersecting poses. Therefore, we use this setup to evaluate the plausibility of

Data	Occ. Leg			Occ. Arm+hand			Occ. Shoulder +Upper Arm		
# frames	60	120	240	60	120	240	60	120	240
Method									
VPoser [PCG*19]	2.53	2.57	2.54	8.51	8.52	8.59	9.98	9.49	9.48
HuMoR [RBH*21]	5.60	6.19	9.09	7.83	8.44	10.25	**4.75**	**5.11**	**4.95**
Pose-NDF	**2.49**	**2.51**	**2.47**	**7.81**	**8.13**	**7.98**	7.63	7.89	6.76

Table 7.3 **Motion estimation from partial 3D observations**: We compare per-vertex error (in *cm*). It can be seen that for leg and arm/hand occlusions, Pose-NDF reconstructs the pose better than VPoser and HuMoR. For occluded shoulders, HuMoR takes the lead. We observe that the results of Pose-NDF depend on the initialization of the occluded joint, as it is expected from manifold projection.

the generated poses. For the occluded leg case, VPoser and our prior-based method perform better. We believe this is because the majority of the poses in both AMASS training and test are upright with nearly straight legs, and hence, VPoser is biased towards these poses. For our method, it highly depends on initialization. Since we have used an initialization close to the rest position, our optimization method generates smaller errors for occluded legs but higher errors for occluded arms and shoulders, as they usually are farther away from the rest pose. For HuMoR, the motion generated is realistic and plausible, but in some cases, it results in a large deviation from the ground truth because the correction in the input pose accumulates over time. We show qualitative results of estimating 3D poses from partial observations. We observe that VPoser is biased towards mean poses (*e.g.* as seen in Figure 7.6), VPoser tends to produce nearly straight legs for occluded leg cases, which is more commonly seen in training data. On the other hand, HuMoR and PoseNDF produce more variety, but in some cases of HuMoR, the resulting pose looks unrealistic due to the accumulation of corrections in the input pose. For our method, it highly depends on initialization. Since we initialise randomly near the mean pose in our experiments, it looks similar to VPoser in most cases.

We compare the runtime of each method used in optimization-based pose recovery methods. For motion denoising and motion recovery from partial observations, we conducted our experiments on a V100 GPU. For image-based reconstruction, we used an RTX 3080 Ti. HuMoR for motion denoising and pose recovery from partial observation takes approximately 10.83 sec. and 10.23 sec., respectively, for one frame. On the other hand, VPoser-based optimization is way faster and takes only 0.96 sec. per frame. Pose-NDF based optimization takes even less: only 0.56 sec for motion denoising and partial observation experiments. This is because Pose-NDF is fast per step and has fewer optimization steps, as compared to VPoser and HuMoR.

7.4.4 3D Pose Estimation from Images

We now show that Pose-NDF can also be used as a prior in optimization-based 3D pose estimation from images [PCG*19]. We use the objective function proposed in SMPLify-

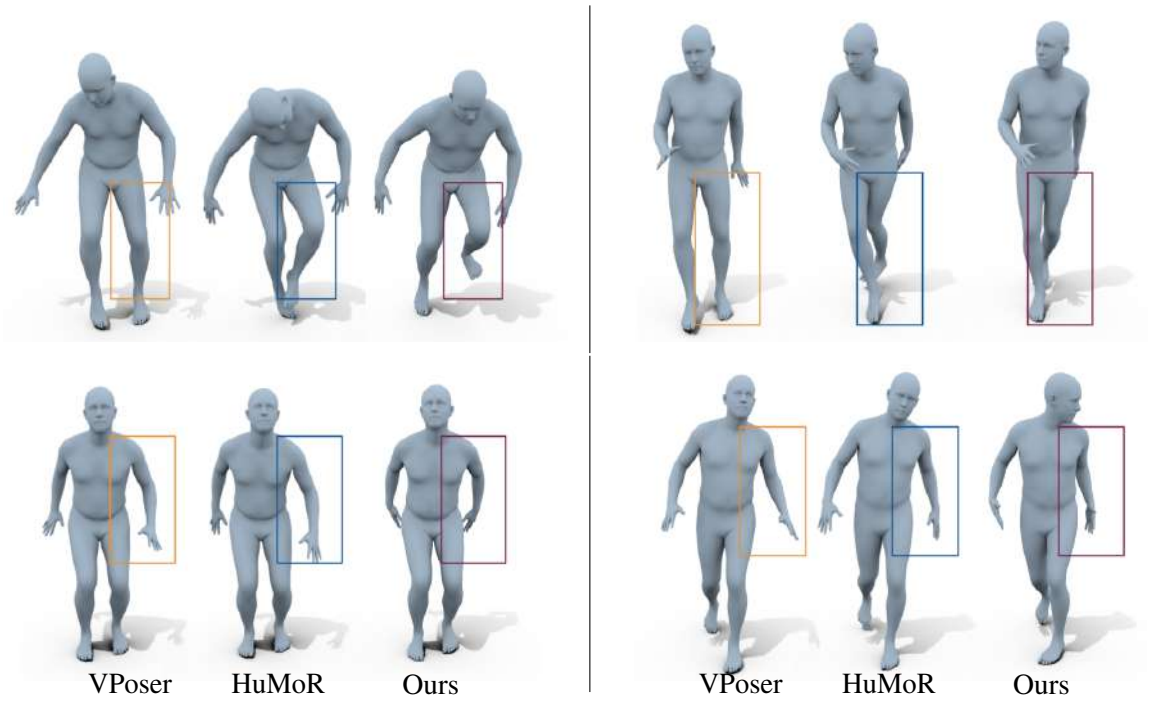


Fig. 7.6 **Fitting to partial observation:** We compare VPoser [PCG*19], HuMoR [RBH*21], and Pose-NDF based prior on the task of recovering 3D pose from partial observation. *Top*: In case of occluded left leg, we observe that VPoser produces nearly straight legs(close to mean position), while HuMoR and PoseNDF produce much more diverse poses. HuMoR results in extremely unrealistic poses in some cases. *Bottom*: We observe similar behavior for the occluded arm case.

X [PCG*19], see Equation (7.13). Since we are working with SMPL body only (without hands or faces), we remove the respective loss and prior terms. Thus, we find the desired pose $\hat{\theta}$ and shape $\hat{\beta}$ as:

$$\hat{\beta}, \hat{\theta} = \arg\min_{\beta,\theta} \mathcal{L}_J + \lambda_\theta \mathcal{L}_\theta + \lambda_\beta \mathcal{L}_\beta + \lambda_\alpha \mathcal{L}_\alpha, \tag{7.13}$$

with data term $\mathcal{L}_J$, bending term $\mathcal{L}_\alpha$, shape regularizer $\mathcal{L}_\beta$, and prior term $\mathcal{L}_\theta$. The data term and the bending term are given as:

$$\mathcal{L}_J = \sum_{i \in \text{joints}} \gamma_i w_i \rho(\Pi_K(R_\theta(J(\beta))) - J_{\text{est,i}}) \tag{7.14}$$

$$\mathcal{L}_\alpha = \sum_{i \in (\text{elbow,knees})} \exp(\theta_i) \tag{7.15}$$

where $J_{\text{est,i}}$ are 2D pose keypoints estimated by a SoTA 2D-pose estimation method [CHS*19], R_θ transforms the joints along the kinematic tree according to the pose θ, Π_K represents a 3D



LSP dataset [JE10]



High resolution LSP dataset [JE10]



COCO dataset [LMB*14a]



3DPW dataset [vMHB*18a]

Fig. 7.7 **Human pose and shape estimation from in-the-wild images:** We use Pose-NDF as a prior in optimization based method to estimate 3D pose and shape from images.

Method	VPoser [PCG*19]	GAN-S [DRC*22]	Pose-NDF
Per-vertex error (*mm*)	60.34	59.18	57.39

Table 7.4 **Optimization-based 3D pose and shape estimation.** Comparison of optimization methods using different priors: VPoser [PCG*19], GAN-S [DRC*22], and Pose-NDF.

Method	ExPose	+No prior	+VPoser [PCG*19]	+GAN-S [DRC*22]	**+Pose-NDF**
Per-vertex error (*mm*)	54.76	99.78	67.23	54.09	53.81

Table 7.5 **ExPose and ExPose + Optimization results.** We evaluate the proposed prior and optimization pipeline to enhance the SoTA 3D pose and shape estimation network, ExPose [CPB*20].

to 2D projection with intrinsic camera parameters and ρ represents a robust Geman-McClure error [GM87]. Further, the bending term $\mathcal{L}_\alpha$ penalizes large bending near the elbow and knee joints, and the shape regularizer is given as $\mathcal{L}_\beta = ||\beta||^2$ [PCG*19]. For VPoser, the prior term is given as $\mathcal{L}_\theta = ||z||_2^2$, where z is the 32-dimensional latent vector of the VAE. In our model, we use $\mathcal{L}_\theta = f^{\text{udf}}(\theta)$ and minimize the distance of the pose from our learned manifold using our projection algorithm. We leverage the distance information provided by our model in optimization by setting $\lambda_\theta = w f^{\text{udf}}(\theta)$, where w has a fixed value. This ensures that if the pose is getting close to the manifold (*i.e.* $f^{\text{udf}}(\theta)$ is very small), the prior term is down-weighted, which results in faster convergence.

We use the EHF dataset [PCG*19] for quantitative evaluation and compare our work with the state-of-the-art priors VPoser [PCG*19] and GAN-S [DRC*22]. A Pose-NDF prior term slightly improves on the VPoser and GAN-S based optimization (Table 7.4). We observe that the neural network-based model ExPose [CPB*20] outperforms all optimization-based results. However, we show that such methods can benefit from an optimization-based refinement step. We refine the ExPose output using Equation (7.13) with Pose-NDF as prior and compare this refinement with no-prior and other priors (Table 7.5). Without prior knowledge, the optimization objective only minimizes the joint projection loss, resulting in unrealistic poses. In contrast, GAN and Pose-NDF improve the result (both qualitatively and quantitatively), generating more realistic poses, while Pose-NDF outperforms the GAN prior. Finally, in Figure 7.7 we show qualitative results of optimization-based 3D pose estimation on in-the-wild images from 3DPW [vMHB*18a], LSP [JE10] and MS-COCO [LMB*14a] datasets.

7.4.5 Pose Generation

We evaluate our model on the task of pose generation. Due to our distance field formulation, we can generate diverse poses by sampling a random point from $SO(3)^K$ and projecting it onto the manifold (Section 7.2.4). We compare the results of our model with sampling

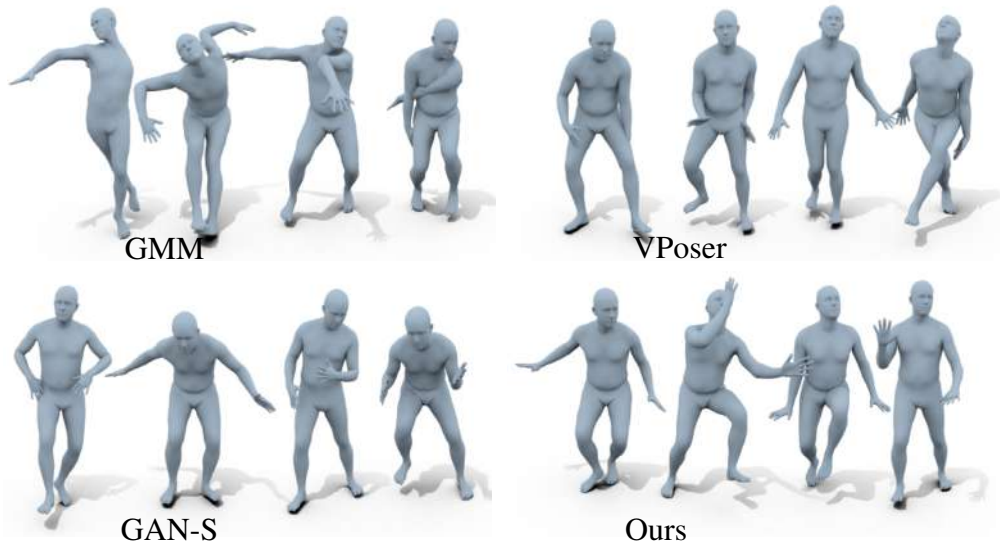


Fig. 7.8 **Diversity and realism in pose generation:** GMM tends to produce highly unrealistic poses, while VPoser, GAN-S, and Pose-NDF generate significantly more realistic ones. According to the APD metric, Pose-NDF exhibits greater pose variance compared to VPoser and GAN-S, leading to a higher diversity in the generated poses.

from the state-of-the-art pose prior VPoser [PCG*19], GMM [BKL*16, OLPM*18], and GAN-S [DRC*22] in Figure 7.8. We use Average Pairwise Distance (APD) [ASSS*20], to quantify the diversity of generated poses. APD is defined as the mean joint distance between all pairs of samples. We also calculate the percentage of self-intersecting faces in generated poses to evaluate one aspect of realism in poses. We randomly sample 500 poses for each GMM, VPoser, GAN-S, and Pose-NDF, and report APD and percentage of self-intersecting faces in Table 7.6. While the GMM yields the highest APD (**48.24** cm), the sampled poses are often unrealistic, as evidenced by visual artifacts in Figure 7.8 (top-left). In contrast, Pose-NDF achieves a good balance—producing more diverse poses than both VPoser and GAN-S, with an APD of **32.31** cm—while maintaining high realism. This is further supported by the percentage of self-intersecting faces, where Pose-NDF outperforms all baselines with the lowest rate of **0.89%**. These results suggest that Pose-NDF is effective at generating diverse yet physically plausible human poses.

7.4.6 Pose Interpolation

Pose-NDF learns a manifold of plausible human poses, so it can be used to interpolate between two distinct poses by traversing the manifold. Specifically, for any given pose, we first project start (θ_0) and end pose (θ_T) on our manifold using Eq (7.7), to get θ'_0 and θ'_T. We then move along the direction of θ'_T from θ'_0 with step size τ using Eq (7.16). The interpolated pose (θ_t) is again projected on the manifold to get a realistic pose (θ'_t). In the subsequent interpolation steps, we move from θ'_t to θ'_T, where θ'_t is updated after each step.

Method	Average Pairwise Distance (cm) ↑	Self-Intersecting Faces (%) ↓
GMM	**48.24**	–
VPoser [PCG*19]	23.13	2.10
GAN-S [DRC*22]	27.52	1.43
Pose-NDF(Ours)	**32.31**	**0.89**

Table 7.6 **Diversity and realism of generated poses from different priors:** We report Average Pairwise Distance (APD) in cm and percentage of self-intersecting mesh faces for 500 randomly sampled poses. Higher APD indicates more diversity, while a lower intersection percentage indicates better realism.



Fig. 7.9 **Comparison with VPoser [PCG*19] on pose interpolation:** We observe that VPoser-based interpolation (*top*) is less smooth than Pose-NDF-based interpolation (*bottom*).

$$\theta_t = \theta'_{t-1} + \tau(\theta'_T - \theta'_{t-1}) \tag{7.16}$$

We compare the results of Pose-NDF with those from VPoser [PCG*19] and GAN-S [DRC*22] interpolation. For VPoser [PCG*19], we project the start and end pose into the latent space and perform linear interpolation using the latent vectors. For GAN-S [DRC*22], we use the spherical interpolation in latent space, as suggested in the work.

We qualitatively evaluate the interpolation quality by calculating the mean per-vertex distance between consecutive frames. A smaller value means smooth interpolation. We compare the interpolation smoothness of different latent space methods using the mean per-vertex distance metric. The Pose-NDF-based and GAN-S interpolations result in similar average distances—2.72 and 2.71, respectively—with relatively low variance, indicating smooth transitions in generated shapes. In contrast, VPoser exhibits a slightly lower mean

	Pose-NDF-based (Ours)	GAN-S [DRC*22]	VPoser [PCG*19]
Mean ± Std (Per-Vertex Distance)	2.72 ± 2.16	2.71 ± 2.45	2.53 ± 4.62

Table 7.7 **Smoothness in pose interpolation:** Mean per-vertex distances (± standard deviation) between consecutive poses. Lower values indicate smoother interpolations.

distance (2.53) but a substantially higher standard deviation (4.62), suggesting less consistent interpolation behavior. These results indicate that both Pose-NDF (represented by Pose-NDF) and GAN-S allow for smooth interpolation, whereas the high variance in VAE-based methods, such as VPoser, suggests that input space distances are not consistently preserved. We tabulate the result in Table 7.7. We compare VPoser based interpolation with Pose-NDF in Figure 7.9) and observe large jumps in VPoser interpolation. This behaviour is not observed in Pose-NDF-based interpolation. Since the VAE learns a compact latent representation of poses, the distance between two input poses is not preserved in the latent space. We provide more results of pose interpolation using Pose-NDF in Figure 7.10.

7.4.7 Pose-NDF vs. Gaussian Assumption Models

Prior work [RBH*21, PCG*19] uses VAE-based models as pose/motion prior, which follow a Gaussian assumption in the latent space. This has three major limitations, as mentioned in Section 7.1. Conversely, Pose-NDF learns the manifold directly in the pose-space without such assumptions and, hence, overcomes these limitations.

Method	$\pm\sigma$	$\pm 2\sigma$	$\pm 3\sigma$
Pose-NDF	**8.18** cm	**8.20** cm	**8.21** cm
VPoser	8.35 cm	9.11 cm	9.13 cm
HuMoR	10.08 cm	11.38 cm	16.86 cm

Table 7.8 **Cumulative error across deviation levels from the mean pose (σ, 2σ, 3σ):** We evaluate per-vertex reconstruction error(in cm) for AMASS Noisy data (60 and 120 frames). We compare the performance of Pose-NDF, VPoser, and HuMoR on motion denoising.

We report the cumulative error based on deviation from the mean pose. We evaluate on AMASS Noisy (60 and 120 frames) and report cumulative error for samples with $\sigma, 2\sigma, 3\sigma$ for both Pose-NDF and VPoser motion denoising in Table 7.8. Pose-NDF demonstrates consistently low error across all deviations, indicating robustness to pose variability. VPoser performs comparably to Pose-NDF for near-mean poses but exhibits increased error with higher deviations. HuMoR's performance degrades significantly as the deviation increases, suggesting its limited ability to generalize to out-of-distribution poses. For both the error increases for samples deviating from the mean pose. And this further highlights the importance of evaluating robustness across a wide range of pose variations.

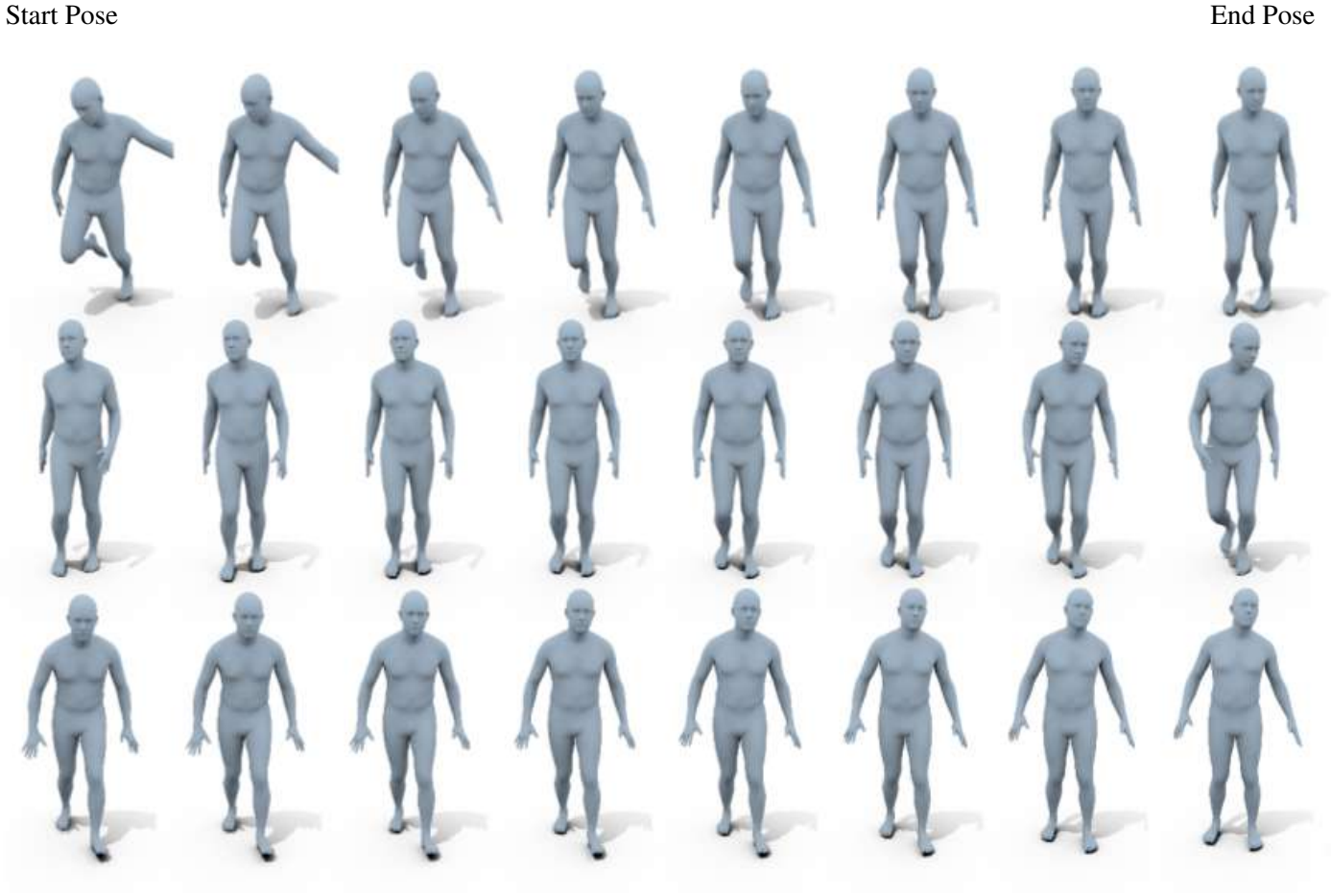


Fig. 7.10 **Pose Interpolation using Pose-NDF:** We observe smooth and more consistent interpolation using Pose-NDF.

Since the Gaussian distribution is mathematically unbounded, it inherently assigns non-zero probability mass to regions far from the observed training data. This leads to the presence of so-called *dead regions* in the latent space—areas where the generative model has not seen any data during training and thus has no support. When sampling from these regions, as is common in Gaussian-based models like GMMs and VPoser, the generated samples often fall outside the manifold of valid human poses. This results in non-plausible, as illustrated in Figure 7.8.

In contrast, our method learns a pose manifold directly from data, ensuring that all sampled poses lie on or near the high-density regions of the learned distribution. This manifold-aware learning not only mitigates the risk of sampling unrealistic poses but also provides more meaningful geometric structure in the latent space. As a result, the distances between points in latent space more accurately reflect pose similarity in the original space. This property is particularly valuable for applications like interpolation and motion synthesis. Specifically, our model enables smooth and consistent interpolations between poses, preserving natural transitions and avoiding discontinuities or artifacts. This is in stark contrast to VPoser, where interpolated poses can drift into poorly represented regions of the latent space, producing unrealistic transitions (see Section 7.4.6).

Inference runtime: We evaluate and compare the runtime performance of various optimization-based pose recovery methods across three tasks: motion denoising, motion recovery from partial observations, and image-based pose and shape reconstruction. To ensure a fair comparison, we conduct all motion-related experiments on an NVIDIA V100 GPU, while image-based reconstruction experiments are carried out on an NVIDIA RTX 3080Ti GPU, given the higher computational demand associated with image processing and feature extraction.

Among the evaluated methods, HuMoR [RBH*21] is the most computationally intensive. For a single frame, it requires approximately 10.83 seconds for motion denoising and 10.23 seconds for pose recovery from partial observations. This high runtime is largely due to HuMoR's recursive structure and the complexity of its optimization process, which involves numerous iterative updates over a high-dimensional parameter space.

In contrast, VPoser-based optimization is significantly more efficient. It completes both motion denoising and partial observation recovery in just 0.96 seconds per frame, thanks to the use of a learned latent pose space that reduces the number of parameters and accelerates convergence. For the image-based pose and shape reconstruction task, VPoser takes 3.63 seconds per frame, reflecting the additional cost of handling high-dimensional visual inputs but still maintaining reasonable efficiency.

Pose-NDF-based optimization further improves runtime performance. It achieves the fastest per-frame runtimes of 0.56 seconds for both motion denoising and recovery from partial observations. Despite handling comparable tasks, Pose-NDF is faster than VPoser due to its lightweight per-step computation and the requirement for fewer optimization iterations. Even for the more complex image-based reconstruction task, Pose-NDF remains efficient, taking 4.59 seconds per frame, which is only marginally slower than VPoser, despite offering improved accuracy in some cases.

In summary, these comparisons highlight the balance between computational cost and accuracy across different methods. While HuMoR provides expressive modeling capabilities for temporal data, it comes at a high runtime cost. VPoser strikes a balance with reasonable efficiency and good performance, and Pose-NDF offers a compelling alternative for scenarios where low latency and scalability are critical.

7.5 Conclusion

We introduced a novel human pose prior model represented by a scalar neural distance field that describes a manifold of plausible poses as a zero-level set in $SO(3)^K$. The method extends the idea of classic 3D shape representation using neural fields to higher dimensions of human poses, mapping quaternion-based poses to an unsigned distance value that represents the distance to the pose manifold. The resulting network can be used to project arbitrary poses to the pose manifold, opening applications in several areas. We comprehensively evaluate the performance of our model in diverse pose sampling, pose estimation from images, and motion denoising. We demonstrate that our model generates poses with significantly greater diversity than prior VAE-based works, achieving state-of-the-art results in image reconstruction and motion estimation.

7.5.1 Limitation and Discussions

On one hand, PoseNDF-based formulation tackles the limitations of Gaussian-based models such as VPoser [PCG*19], HuMoR [RBH*21]. For more details, please check Section 7.4.7. Training a good pose manifold comes with its own challenges. The distribution of training samples plays a significant role in shaping the learned pose manifold. Prior work on neural implicit representations [CMPM20, TSTPM21] for 3D shape modeling has demonstrated that concentrating samples near the object surface leads to significantly improved reconstruction fidelity. Based on this, we also introduce a heuristic sampling scheme to preferentially sample more training data points corresponding to plausible and low-noise human poses during training, thus facilitating robust and fine-grained manifold learning. However, later work NRDF [HTB*24] observes that developing such a heuristic is challenging for high-dimensional data such as human poses. NRDF systematically investigates the influence of training data distribution on the learning of distance field priors for pose manifolds. To enhance sample density in proximity to the surface while ensuring a smooth fall-off with increasing geodesic distance, NRDF introduces a principled sampling strategy grounded in Riemannian geometry.

7.5.2 Future Work and Impact

Follow-up works have extended the formulation introduced in PoseNDF to improve the learned human pose prior or extended it to model the human motion manifold. For example, NRDF [HTB*24] utilized NDFs for manifold learning, specifically leveraging Riemannian distance metrics to effectively capture intrinsic geometric structures. Furthermore, Gan *et*

al. [GNF*25] explored the application of NDFs in constructing informative 2D pose priors expressed in polar coordinates, thus providing richer geometric cues to guide pose estimation tasks.

Momanifold [DFZ*24] introduced robust motion models that exploit manifold representations learned via NDFs. MoManifold is a novel human motion prior, which models plausible human motion in continuous high-dimensional motion space. This makes human dynamics explicitly quantified to a score and thus can measure human motion plausibility. Specifically, Momanifold proposes novel decoupled joint acceleration manifolds to model human dynamics from existing limited motion data. Recent work, such as [RLGCC23], also extends Neural Implicit Functions for manifold learning.

8

Conclusion

A human presence is defined by how we look(*e.g.* facial features, hairstyle) [YTB*21, LOZ*24, EST*20, KZB*24], how we behave(*e.g.* motion, pose) [GZW*20, TRG*23] and how we dress(clothing texture, size, dynamics) [LCT18b, TBTPM20, GTBH23] and how we interact with world(*e.g.*, with objects or other people) [BXP*22, JZL*24, MYP*24]. In this thesis, we have explored mainly how we dress [TSTPM21, TBTPM20] and how we behave statically [TAL*22]. In this chapter, we provide concluding remarks on these contributions and outline future directions for research in digital humans. We position our work within the broader landscape of digital human modeling and understanding, highlighting how our findings contribute to the overall goal of building rich, controllable, and realistic virtual human representations. We also discuss the practical relevance of this research for real-world applications, including virtual try-on systems, gaming, telepresence, content creation, and immersive digital experiences.

In Part I of this thesis, we advance the field of digital human modeling by moving from traditional template-based mesh representations to neural implicit-based digital avatars, ultimately aiming for more foundational and unified models of humans in clothing. In Chapter 4, we propose the first template mesh-based representations to learn a complex and subtle size-dependent clothing deformation model. Building on this, we transition to more dynamic, pose-dependent deformations for complex garments using neural implicit representations in Chapter 5, where we reinterpret classical concepts such as skinning as an implicit skinning field and displacement maps as continuous deformation fields. These chapters also introduce datasets designed to study both types of deformations. Our models and datasets, particularly SIZER and the ClothSeq dataset, have been widely adopted in subsequent research for a variety of human understanding tasks, such as image-based 3D reconstruction [CPA*21a], clothing resizing, scan fitting [DLLG*23, CJS*22], and animation [QXL*22].

In Chapter 4, we also introduce layered representations of clothed humans. Specifically, we introduce a parsing network that transforms single-layer human meshes—characterized by coherent topology—into semantically meaningful, multi-layered representations. These layered models enable several downstream applications, including cloth retargeting, editing tasks such as clothing resizing and draping, and enhanced semantic understanding of garments. Since the real-world human clothing is inherently semantic, layered, and disentangled, we revisit the modeling paradigm in which the human body and clothing are treated as distinct layers. Both layers, however, are driven by a shared underlying body model. Our goal is to scale this idea to generalize across a wide range of clothing types—something that has remained a significant limitation in prior works [LGF23, HYZ24, DLLG*23].To address this, we introduce a large-scale, unified model of clothing in Chapter 6, which we

refer to as TailorDiff. Unlike traditional mesh-based approaches, TailorDiff is built on a point cloud-based representation, making it topologically agnostic and capable of capturing diverse garment styles within a single unified framework. Moreover, this model is compatible with the widely-used SMPL body model. TailorDiff is a step forward in representing clothed humans in a more physically realistic and semantically rich manner. It can serve as a substitute for SMPL in scenarios involving clothed humans, and can be applied to tasks such as image-based 3D reconstruction, scan fitting, clothing editing, and reposing. Moreover, due to the explicit nature of the model, it enables the generation of layered representations of clothed humans, which are crucial for high-level tasks involving physically realistic clothing interaction with both the body and clothing. This progress—from class-specific and task-limited models to a unified, foundational model—opens up exciting new directions in digital human modeling. This work marks a substantial step forward in understanding humans in clothing and lays a strong foundation for future developments in areas such as controllability, animation, content generation, editing, and other tasks in computer vision and computer graphics.

In Part II of this thesis, we explored the fundamental question of how humans behave and move through the lens of body poses. Parametric body models such as SMPL [LMR*15] offer a compact and controllable representation of digital humans by mapping human body shapes and poses into a low-dimensional parameter space. However, these models often lack meaningful constraints on their pose space, which can lead to unrealistic and physically implausible poses. To mitigate this, pose priors are introduced to regularize the space of plausible human poses.

In Chapter 7, we introduced a novel formulation for learning a pose prior, which is based on high-dimensional neural distance fields. Instead of learning a probability density under the assumption of some tractable and well-defined prior, PoseNDF [TAL*22] learns how far a pose is from plausibility using a neural distance field. As a result, this formulation avoids the bias introduced by assuming a specific form of the prior, which is commonly Gaussian [PCG*19, BKL*16]. The distance field provides gradients to project noisy data points onto the manifold and is used as a pose prior for downstream tasks. Beyond its practical utility, PoseNDF opens up a broader perspective on learning data manifolds using neural implicit representations. Several works employ neural implicit or neural distance field-based manifold learning for motion [DFZ*24] or human poses [HTB*24, GNF*25]. A recent follow-up [RLGCC23] extends this paradigm beyond articulated human bodies to image data, highlighting the generality of neural implicit-based manifold learning. In summary, this chapter demonstrates that reframing pose priors through neural distance fields not only improves existing tasks in pose estimation but also lays the groundwork for future research into learning complex, high-dimensional data manifolds using implicit neural representations.

In the following sections of this chapter, we begin by discussing the key insights derived from each chapter of the thesis in Section 8.1. This section distills the key findings and highlights the overarching themes that reflect our collective understanding and learning from the different parts of the thesis. Next, in Section 8.2, we outline several promising research directions and ideas that this thesis paves the way for. These are high-level suggestions aimed at the broader research community. For future work that is specific to individual

contributions, we encourage readers to refer to the respective chapters where these ideas are discussed in more detail.

Given that research in digital humans can be highly privacy-sensitive and raises important ethical concerns, we devote Section 8.3 to exploring the societal implications of our work. Finally, in Section 8.4, we reflect on the implications of recent advancements in large-scale foundational models and the shift towards more generalist approaches in computer vision. We offer our perspective on how this broader context influences the direction of digital human research and how our work fits within this evolving landscape.

8.1 Key Insights

8.1.1 Human in Clothing

In this section, we first discuss the key insights from each work presented in this thesis and then conclude with a unifying summary for all chapters in Part I of the thesis.

8.1.1.1 SIZER: Parsing and Learning Size Sensitive 3D Clothing

We first discuss the key insights specific to the two models, namely *ParserNet* and *SizerNet*, as presented in Chapter 4 of this thesis.

Interpretable and Detail-Preserving Clothing Parsing from 3D scans: We introduce a novel method for parsing clothing from 3D scans. Instead of directly regressing a dense displacement field using a fully connected MLP, our approach predicts a sparse matrix that linearly combines input mesh vertices to reconstruct each clothing vertex. This formulation offers both locality and interpretability: each row in the sparse weight matrix explicitly indicates which input vertices influence a given clothing vertex, transforming an otherwise black-box network into an explicit, geometry-aware mapping. Moreover, since our method copies geometry already present in the scan rather than synthesizing it, fine details are preserved by design. In contrast, an MLP suffers from mean-square bias: with an L2 reconstruction loss, the global optimum for each vertex coordinate becomes the conditional mean of that coordinate over the training set. As a result, high-variance, high-frequency details are averaged out—since they cancel across samples—while low-frequency structure, being consistent, is retained.

This simple shift (i) embeds geometric priors into the network, (ii) yields interpretable weights, and (iii) improves accuracy and visual fidelity, especially on challenging garments such as very loose clothing or those with fine details.

Size-aware Draping Network learned from real 3-D scans: In SIZER [TBTPM20], we introduce the first size-aware garment-draping network, learned entirely from real-world 3D scans. Training SizerNet requires an expensive pipeline of data capture, reconstruction, and garment–body registration, but the effort pays off: the model learns the subtle, non-linear effects of resizing that simple uniform scaling misses. Our encoder–decoder architecture

conditions on both the discrete size label and continuous SMPL body shape coefficients, letting it modulate looseness and tightness realistically around sleeves, waistbands, and hems—areas where naive linear scaling fails. By bridging the gap between nominal size labels and true 3-D fit, SizerNet provides a fast, data-driven alternative to slow physics simulation, enabling practical virtual try-on and avatar pipelines.

8.1.1.2 NeuralGIF: Neural Generalized Implicit Functions for Animating People in Clothing

In subsequent paragraphs, we discuss the key insights from Chapter 5 regarding rigging and learning pose-dependent displacement for animating digital humans using neural implicit-based representation.

Beyond Templates with Neural Implicit Avatars: Traditional mesh-based avatar pipelines rely on expensive and brittle scan-to-template registration and are ultimately constrained by the fixed resolution and topology of the mesh. By switching to neural implicit representations, we eliminate these costly pre-processing steps and instead learn directly from scans. To make this possible, we reinterpret classical mesh concepts—such as skinning weights, pose, and shape-blend shapes—as continuous volumetric fields parameterized by neural networks. We provide a detailed formulation and comparison in Section 5.4.

Consequently, we introduce backward-skinning, a learnable field that maps any point in a posed frame back to canonical space, extending SMPL's linear-blend skinning (LBS). Additionally, we train a pose-conditioned displacement field (and an optional shape-conditioned term for multi-shape datasets) that operates similarly to SMPL's blend shapes, but without being limited to mesh vertices. Representing both the skinning weights and the displacement volumetrically enables the model to capture loose clothing that deviates significantly from the underlying body, while the learned skinning field accommodates complex, highly non-linear pose-dependent deformations.

Pose-dependent Backward Skinning Field: In [TSTPM21], we introduce a volumetric backward-skinning field that learns highly complex, pose-dependent deformations of clothing and/or skin directly from raw scans. Unlike mesh-based avatars, which rely on predefined linear-blend skinning with a fixed template, realistic garments need far more degrees of freedom. Because NeuralGIF (and similar neural-implicit methods) learn directly from posed scans and never observe a canonical rest shape in multiple poses in their training data, a forward-skinning formulation is ill-posed: a single point in posed space may correspond to several canonical locations, forcing some works to rely on slow, iterative root-finding. Our backward-skinning formulation resolves this by learning, in one forward pass, a mapping from posed space back to canonical space—no iterative solver required—while simultaneously learning a rich, pose-conditioned skinning field. The result is a model that retains high geometric fidelity even under large cloth-to-body deviations and supports fast, fully differentiable evaluation. A known limitation is generalization to extreme poses, because the skinning field is trained on a finite pose distribution, performance can degrade on extreme, unseen articulations.

8.1.1.3 TailorDiff: A Unified and Expressive Generative Model for 3D Clothing

We describe *TailorDiff* as an SMPL-like *foundation model* for clothing. It unifies garment categories, supports layered editing, and remains fully differentiable—properties that earlier methods lack. In subsequent paragraphs, we discuss the key insights from Chapter 6.

A Foundational Clothing Model: TailorDiff plays the same role for dressed humans that SMPL plays for naked body shape: a single, category-agnostic prior able to synthesise T-shirts, skirts, jackets, trousers, indeed clothing of *any* topology—within one network. By explicitly disentangling body and cloth, the model can be directly used in graphics and vision pipelines for generation, editing, or as a prior, eliminating the need for separate models per clothing class. This generality is made possible by (i) a point-cloud representation that naturally handles varying topology, and (ii) a large-scale, heterogeneous training data assembled from multiple public datasets. Together, this enables the first practical substitute for SMPL in clothed-human modelling, offering an explicit, layer-aware, and topology-flexible alternative that can be fitted to scans and recover parametric, layered descriptions of people in clothing.

Fine-grain Controllability due to Hierarchical Latent Space: TailorDiff introduces a two-level latent architecture. A *global latent vector* for coarse shape and a *per-point latent cloud*, along with a learned occupancy of clothing on the body that refines local details such as sleeve length, hemlines, and other subtle details. Latent interpolation, partial denoising, and direct latent editing therefore provide users and optimization-based methods with precise, continuous control that earlier approaches could not match.

Solving Inverse Problems: Beyond generation, TailorDiff serves as a prior for downstream tasks such as scan fitting. Given a 3D scan, a latent-code optimisation recovers full-layered garments—including those occluded in the scan—while preserving the correct cloth-to-body ordering. Because of the unified model, TailorDiff latents can explain any type of clothing, as opposed to prior works that are limited to fixed clothing. This makes the approach practical for applications such as virtual try-on, digital wardrobe capture, and scan completion. However, since we use VAE latent space for these tasks, our model is limited to good initialisation of latents; otherwise, this might lead to being stuck at bad minima.

Explicit Body Encoding: The explicit coarse body encoding as conditioning results in better body and clothing alignment. Removing this cue increases cloth-to-body intersections and noticeably wrong alignment. Moreover, since the model takes an explicit point cloud of the underlying body as input, the method is gender and body model-agnostic, which was not possible for prior and concurrent works, and is mostly biased towards female body clothing models.

In the following paragraphs, we discuss the overarching theme that spans all chapters in Part I of the thesis, with a primary focus on the choice of representation for each task.

3D Representation for Humans in Clothing: Across the three chapters in Part I, we explore different 3D representations and outline advantages and disadvantages of each

representation concerning applications and ease of learning and modeling. In brief, we relax geometric assumptions, transitioning from fixed-topology meshes to topology-free neural implicit surfaces, and ultimately to topology-free point clouds. In Chapter 4, the body is represented with SMPL+D while each clothing is represented as a SMPL+G sub-mesh. The explicit vertex connectivity eases learning, allows classical regularisers such as the Laplacian, and plugs straight into standard graphics/animation pipelines. However, this is not scalable. Every new clothing category requires a new template and a dedicated network, an expensive and error-prone pre-processing step that fails on very loose garments.

Chapter 5 therefore abandons templates and learns a signed-distance field whose backward volumetric skinning field maps posed points to canonical space. This captures arbitrary topology and high-frequency wrinkles and learns directly from a scan, but inference is slower and expensive. Since the representation assumes watertight surfaces, modeling clothing and layered clothing remains challenging to model using this formulation.

Finally, Chapter 6 returns to an explicit but unordered point-cloud representation. Points are inherently topology-agnostic, render efficiently via modern splatting, and—when combined with a hierarchical VAE and latent diffusion—form a single generative prior that can represent all diverse kinds of clothing using a single model. This has not been done before. However, this comes with surface ambiguity: without an explicit manifold, the output can be noisy, and downstream tasks, such as physics-based simulation or meshing, require an extra processing stage.

In short, the trajectory moves from template-bound accuracy to detail-rich but computationally intensive implicits, and then to scalable, category-agnostic point clouds. Each representation trades structural assumptions for representation power, and the best choice depends on the downstream tasks. A fully unified model remains an open problem, but a hybrid point-cloud + distance-field design that incorporates surface awareness with topological freedom is a promising next step.

8.1.2 Human Pose Priors

In this section, we first discuss the key insights from Chapter 7 of Part II of the thesis, where we focus mainly on the advantages of our novel distance field-based formulation over Gaussian-based assumptions and discussion around latent space and training data for learning pose prior.

Gaussian assumptions v/s Neural Distance Fields: Most prior works for learning pose priors have Gaussian assumptions either in the latent space or input pose space. This results in some limitations due to Gaussian assumptions, which our neural distance field formulation overcomes. Gaussian-based methods have a bias toward the mean and favor “average” poses, underrepresenting rare but still valid poses. Since NDF-based prior treats the whole manifold uniformly, infrequent but correct poses are treated on an equal footing with common and average poses. Gaussian-based methods have dead regions in latent space. Folding the pose manifold into a single Gaussian leaves large “holes” where sampling produces implausible poses. By staying in the original pose space and learning a distance field, PoseNDF keeps

the manifold's true geometry intact, eliminating dead zones.

Latent space v/s Input space: Moreover, we observe that small steps in latent space can translate to large and unrealistic jumps in final poses, resulting in non-smooth interpolation. Because the field is trained with a geodesic metric on quaternions, Euclidean changes in NDF input correlate with meaningful pose changes, yielding smoother interpolation and more stable optimisation. However, as noted by recent diffusion-based methods [RBL*21], it might be more beneficial to learn a more meaningful latent space, before learning the distance field.

Impact of Training Data Samples: We observe that the sampling strategy of the training dataset has a significant impact on the learned manifold. Prior work on neural implicit methods for 3D shape learning has also shown that sampling more points near the surface yields more detailed reconstructions. PoseNDF heuristically samples more plausible and less noisy poses during training. The follow-up work, NRDF [HTB*24], proposes that in order to learn well-defined and detailed pose manifolds, we must diligently study the role of training data distribution in learning distance field priors. To draw more samples near the surface with a gradual decrease for farther regions, NRDF introduces a well-defined wrapped sampling algorithm on Riemannian manifolds that allows explicit control over the resulting distance distribution.

8.2 Future Direction

In this section, we outline key research ideas and broader directions that emerge from this thesis. For future work related to specific chapters, please refer to the respective chapter conclusions. Here, we focus on overarching goals and long-term opportunities for extending this research.

Hybrid Representations: Points, Distance Fields, and Sewing Patterns

In Part I, we primarily focused on exploring different representations for humans and clothing. Realistic digital humans may ultimately require a hybrid representation that combines the strengths of multiple modeling paradigms. This thesis explored purely implicit neural fields (e.g., NeuralGIF) and point clouds (e.g., TailorDiff). However, future models could leverage joint combinations of point clouds, implicit fields, and even 2D pattern information. For example, one could use a set of 3D point primitives to capture fine details or localized features, while an implicit signed distance field defines the continuous surface that the points lie on. This approach merges the fidelity of explicit representations with the topological flexibility and surface awareness of implicit ones.

Beyond 3D geometry, incorporating garment sewing patterns into the representation offers another dimension of control. With recent advances in large-scale paired datasets of 2D sewing patterns and 3D clothing [KKK*24], a promising direction: learning a joint model of 3D clothing and its 2D patterns, or even learning a model directly from 2D

patterns. In the past year, several works have already modeled clothing using sewing patterns [LYW25, NAK*25, BXX*25, LGF23]. However, converting 2D patterns to 3D remains largely limited by physics simulations. A differentiable, robust, and stable mapping from 2D patterns to 3D would enable far more practical applications, as it could be used for understanding both 2D and 3D data. This demands a hybrid parametric model: the 2D parameterization makes it easier to enforce physical constraints (such as avoiding penetrations and ensuring proper seam alignment) and to support user edits via 2D pattern manipulation, while the 3D mapping captures the complex draped shape and enables closing the loop between discrete patterns and 2D/3D visual observations.

Looking forward, we envision a representation that fuses point-based structures (or learned 3D primitives such as Gaussians), implicit distance fields, and possibly explicit panel definitions. By combining these elements, we can build a model that is flexible (not restricted to a fixed topology), detailed, controllable, and physically grounded, enabling the creation of more realistic and controllable digital humans in clothing than any single representation could achieve independently.

Super Skeletons and Advanced Rigging for Clothing and Accessories

Works presented in this thesis and concurrent works have relied on SMPL rig for animating people in clothing. In this thesis, methods such as NeuralGIF, TailorDiff animated clothing via the standard SMPL body skeleton, meaning clothing was bound rigidly to the body's movements. This simplifies animation but fails to capture independent motions of loose or layered items for complex poses – for instance, a skirt or coat tends to move unrealistically if they are treated as skin-tight body parts. This trend has been followed by almost all concurrent works [LZWL24, WCS*22, SYMB21, CZB*21, CJS*22, MYR*20, PLPM20, GTBH23, SOC22]. One clear avenue for progress is overcoming the rigging limitations of current human models like SMPL. Very few, yet some works exists, which deviate from predfined body rig, *e.g.*, [LWP*23] models learnable skinning with latent bones, NeuroSkinning [LZT*19] predicts skinning weights given a super skeleton, while other works predict the skeleton rig for given 3D/4D scans [XZK*20, ZPG*25] or from videos [LGHT24]. A promising direction is to design "super skeletons" that augment the base human rig with additional bones or deformers for secondary elements like clothing and hair. By explicitly animating these layers (rather than skin-binding them to the body alone), each clothing can sway, fold, and collide naturally with the body and other clothing. Recent work [SCD*25], a diffusion-driven rigging framework, introduces skeletons for hair and clothing alongside the body, yielding far more realistic animations than body-only rigs. Such advanced rigging frameworks allow heterogeneous structures to be controlled, enabling, for example, a ponytail and a flowing dress to react independently to character motion. However, the results are still unrealistic and a bit constrained. Super skeletons would thus push beyond SMPL's single-layer paradigm and equip digital humans with the kinematic richness needed to animate complex, multi-layer outfits in a believable way.

Learning from Monocular Data and Leveraging 2D priors

In this thesis, all the works presented learn directly from 3D or 4D data. Our work has primarily used 3D/4D scan datasets (e.g. subjects in a lab scanner or motion capture suit) to learn models like SIZER, NeuralGIF, and PoseNDF. While scans offer ground-truth geometry, they cover limited poses and outfits due to the cost and complexity of capture. A key forward direction is to leverage recent foundation models and self-supervised learning to train on more ubiquitous sources [KBM*24]: monocular videos, single-view images, and other "in-the-wild" observations where 3D ground truth is absent. Recent motion priors either rely on curated mocap datasets [GZZ*22] or 3D motion estimated from videos [LZL*23, ZLZ*25]. Advancements in large-scale vision models are making this possible. We see the future of 3D digital humans will exploit the vast unlabeled 2D data available, rather than relying solely on controlled 3D scans. For instance, diffusion-based models and vision–language models have been used to inject strong prior knowledge into 3D human reconstruction pipelines [XXMPM24, KAC*24, TDK*24], so that high-fidelity avatars can be inferred from sparse inputs like a single photograph or a text description. A similar trend has also been seen in learning motion prior via 2D diffusion [LLW25]. We can envision training regimes where internet videos supply endless, varied motions and outfits, and the model, guided by deep priors (for example, learned human shape spaces or powerful depth estimators), learns to infer 3D humans under all sorts of conditions. In contrast to the carefully controlled 3D scans, these in-the-wild learning methods will yield models that are more diverse and general. They represent an important step toward truly universally applicable 3D digital humans – avatars that can be built from a smartphone video and deployed in any AR/VR or animation scenario, bringing the vision of realistic digital humans to the masses. However, injecting robust control and robust 3D consistency in these models is still an open and challenging direction.

Physically Plausible and Layered Clothing

A significant next step toward physically plausible clothing is to explicitly model multiple clothing layers and their interactions. Existing methods, such as NeuralGIF, focus on a single layer (e.g., outerwear on a body), ignoring interactions with underlying layers. SIZER and TailorDiff model layers separately but handle inter-layer collisions only through simple heuristics to avoid intersections. However, real-world dressing involves complex layering, such as shirts under jackets or pants over socks, where contact, friction, and collisions influence garment drape. Modeling these effects remains challenging, as multi-layer interactions are hard to learn in neural frameworks [LKL23, SOTC22].

One promising approach is to learn cloth–cloth interactions explicitly. For instance, ClothCombo [LKL23] uses a graph neural network to model inter-garment forces and collisions. It fits each garment independently and then uses an untangling network to resolve interpenetrations, effectively learning inter-layer physics. Another direction leverages 2D panel-based representations [LGF23], where overlap detection and layering order are more intuitive, helping to prevent intersections. Classical physics simulators naturally handle such interactions and produce realistic results; however, they are often computationally expensive,

unstable in the presence of extreme collisions, and non-differentiable. A promising frontier is to combine both worlds: using differentiable physics or neural approximations to merge learning-based models with physical realism. Recent neural simulators [KHTG24, BME22], for example, can now approximate cloth dynamics, including for loose garments, in an efficient and differentiable manner.

Unified Foundational Models for Holistic 3D Humans

Looking further ahead, we foresee the emergence of unified foundational models that encompass the full complexity of digital humans – from body pose to facial expressions, hair, and clothing – within a single coherent framework. Currently, the modeling of humans is fragmented. For example, this thesis addressed body shape and clothing deformation (SIZER [TBTPM20], NeuralGIF [TSTPM21], TailorDiff) and body pose models (PoseNDF [TAL*22]) separately from finer details like facial features, hairstyles, or accessories. Even state-of-the-art parametric models were initially compartmentalized (e.g., SMPL [LMR*15] for the body, FLAME [LBB*17] for the face, etc.). A future foundational model would instead learn a joint representation of the entire human or provide a unification of existing separate models, serving as a one-stop foundation that can be adapted or fine-tuned for various tasks. Such a model could be viewed as the 3D human analog of a language or image foundation model —a general-purpose neural avatar that, once learned, can be utilized for a range of downstream applications. For instance, the same model could be fine-tuned to drive a virtual try-on system (focused on clothing) or a digital actor for movies (focused on facial performance and motion), without requiring retraining from scratch. By building on the integrated approaches initiated in our work (e.g., combining learned clothing deformation with body pose), researchers can aim to consolidate all these elements into a single, controllable 3D human model. The impact would be profound: it would ensure consistency across all elements and likely enable new capabilities, such as joint reasoning about pose, expression, and cloth dynamics when generating or analyzing virtual humans.

Neural Implicits for Generic Data Manifold

In this thesis, we introduced PoseNDF[TAL*22], a novel framework for learning the manifold of plausible human poses using Neural Distance Fields (NDF). As reviewed in Section 3.2.2, several follow-up works have explored NDF-based formulations for modeling data manifolds, particularly in the context of motion [DFZ*24] and pose [HTB*24, GNF*25]. An especially compelling extension is presented by EBIM [RLGCC23], which formulates Energy-Based Implicit Manifolds by representing complex data topologies as the zero-level sets of neural networks.

This opens an interesting direction of using distance- and energy-based methods for learning general data manifolds. It opens promising avenues for applying such formulations to generative modeling in image and video domains. Notably, in Section 7.3.2, we observe structural parallels between sampling procedures in NDF-based models and diffusion models, raising the question of whether NDFs might offer empirical or theoretical advantages over diffusion-based approaches. Furthermore, several open questions remain: Should modeling

be conducted in the latent space or directly in the input space? What is the role of accurate density estimation in downstream generative tasks? Addressing these will be crucial for advancing NDF-based generative modeling.

8.3 Social Implications

Research on controllable 3-D digital humans promises concrete public benefits. High-fidelity avatars can enrich telepresence, distance learning, and remote healthcare; realistic virtual try-on can curb the billions lost yearly to clothing returns and make fashion more inclusive by allowing previews of fit on a wider range of bodies; and provide more controllable digital human and avatar generation and editing. By releasing robust priors for pose and clothing, the thesis also lowers the barrier for creatives and smaller companies to produce accessible AR/VR experiences.

Yet the same realism that enables these positives introduces new risks. Detailed scans can expose sensitive biometric cues (e.g., weight, health indicators, cultural dress) that, if re-identified, threaten privacy and bodily autonomy. Hyper-real neural avatars combined with text-to-video models could be exploited for deep-fake harassment or political misinformation unless provenance metadata and watermarking accompany every release.

Finally, we observe that recent large-scale datasets exhibit demographic bias, and models trained on demographically narrow data may reinforce stereotypes or systematically misfit underrepresented groups.

8.4 A Broader View on Topic

Modern 3-D avatars are represented either as explicit meshes that are skinned and rendered in real time, or as neural implicit fields such as NeRFs or Neural Avatars that store geometry and appearance in a continuous MLP, and point or Gaussian-based hybrids that trade some realism for fast splatting while keeping topology flexible. Meshes provide perfect control, implicit surfaces offer unmatched detail, and point clouds sit between—this triad defines the technical landscape for controllable 3D humans. We have discussed the key insights about these models and possible future direction in Section 8.1 and Section 8.2. Here, we provide a discussion on a broader and orthogonal view on modelling digital avatars, which uses 2D priors and contrasts this with 3D digital avatars.

While 3-D pipelines have dominated games, VR, and VFX, the last year has seen a surge of 2-D diffusion models that directly synthesize videos and images with more realism and diversity [WAW*25, KTZ*24]. These models inherit their visual prowess from billions of training data points and are gradually moving towards more controllable models.

Technically, such models excel at creative breadth and pixel-level realism: they incorporate massive priors over lighting, mood, and cinematography, so prompts ranging from "film-noir detective walking in neon rain" to "low-poly llama on Mars" are fulfilled without ever touching 3-D assets. They also eliminate traditional pre-production steps such as scanning, re-topology, or rigging. These are labor-intensive and domain expert tasks. Some

works have introduced control over camera [ZGV*25], controlled generation given image and skeleton motion [MYCL24, HGZ*23] or virtual-try on [LSZ*24]. Yet without an explicit surface or control signals for a given task, spatial coherence can be impacted. For example, in the case of digital avatars or character animation, hands merge, limbs swap sides, and clothing changes frame by frame, especially during long motions or camera orbits. Direct interactivity is impossible: re-using the exact same character identity across scenes also remains challenging.

These gaps in the direction of controllable digital avatars are steering towards a convergence. Research works like DreamAvatar [CCH*23] optimize a NeRF and pre-trained text-to-image diffusion models for providing 2D self-supervision, effectively bootstrapping a 3D asset from 2D priors. Some recent work leverages multi-view diffusion models to directly predict 3DGS of human avatars from single images [XXMPM24, TCC*24], along with skeleton and skinning weights [SCD*25] for character animation. 3D representations handle collision, physics, real-time lighting, and better control, while 2-D diffusion handles realistic generation, style exploration, and final-pass visuals for non-interactive media. We see a mixture of both these approaches already infiltrating industry and applications. Games and XR will keep explicit rigs for responsiveness but may texture or relight them with image generators; digital-fashion houses can simulate clothing fit and dynamics and then spin marketing videos through video models [Ras25]; movie pre-visualisation can already stitch video model-generated clips into storyboards [Run24] before asset teams recreate key shots in full 3-D. In short, 2-D diffusion has democratised visual creation, but so far only explicit 3-D avatars deliver the persistent geometry and control that interactive worlds demand. The most exciting progress now lies in hybrid pipelines, which provide easy control, rather than requiring expert knowledge, and have explicit 3D representation or understanding to solve physical interactions, marrying the best of both paradigms for the next generation of digital humans.

List of figures

1.1 **SIZER:** A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing. . . . 4
1.2 **NeuralGIF:** Animating people in complex clothing. . . . 5
1.3 **TailorDiff:** A topologically agnostic model for clothing generation, editing and inverse problems. . . . 6
1.4 **PoseNDF:** A NDF-based prior for modeling plausible human poses. . . . 7

2.1 **Linear Blend Skinning** [JDKL14]: *Left*: Bone transformations for a given deformed pose. *Right*: Blend weights corresponding to the lower and upper arm bones. . . . 12
2.2 **SMPL body model:** Template mesh with predefined blend weights, where shape and pose-dependent deformation are added(*middle*) and finally articulated to the given pose(*right*). (Courtesy of [LMR*15]) . . . 14
2.3 **Diffusion Models:** *Top:* Introducing noise via Forward diffusion process. *Bottom:* Modeling the reverse process with a neural network. . . . 18

4.1 **SIZER**: A dataset and model of people with clothing size variation. *Left*: We present the first dataset of 3D scans of people captured in different clothing **styles and sizes**. *Right*: We further introduce clothing parsing and resizing in 3D. Specifically, given a registered scan of a subject, **ParserNet** can segment the upper and lower clothing in 3D, and **SizerNet** can resize the parsed clothing to a desired size. Here, we present the ground truth 3D scan with texture, along with the parsed and resized clothing behind the scan, which illustrates the subtle differences introduced by changes in clothing size when draped on the same body. . . . 41
4.2 **Clothing Parsing and Resizing Model:** We propose a model to estimate and visualize the dressing effect of clothing conditioned on body shape and clothing size parameters. For this we introduce *ParserNet* (f_w^U, f_w^L, f_w^B), which takes a SMPL registered mesh $M(\theta, \beta, \mathbf{D})$ as input and predicts the SMPL parameters (θ, β), parsed 3D clothing using predefined templates $T^g(\beta, \theta, \mathbf{0})$ and predicts body shape under clothing while preserving the personal details of the subject. We also propose *SizerNet*, an encoder-decoder $(f_w^{\text{enc}}, f_w^{\text{dec}})$ based network, that resizes the clothing given as input with the desired size label $(\delta_{\text{in}}, \delta_{\text{out}})$ and drapes it on the body shape under clothing. 45
4.3 **Dataset capture setup:** Sizer dataset scans were captured using Treddy static scanner [tre] with 130 fixed cameras. We show some random camera frames from the capture setup. . . . 48

4.4 **3D Clothing Parsing from single mesh:**(*Left to right*) Input single mesh ($\mathbf{M}^k$), clothing template ($T^g(\beta,\theta,\mathbf{0}) = \mathbf{I}^g T(\beta,\theta,\mathbf{0})$), clothing mesh extracted using $\mathbf{G}^{g,k} = \mathbf{I}^g\mathbf{M}^k$, multi-layer meshes ($\mathbf{G}^{g,k}$) registered to SMPL+G, all with clothing class specific edge connectivity $\mathbf{E}^g$, and segmented scan $\mathbf{G}^{g,k}_{\text{seg}}$ with instance specific edge connectivity $\mathbf{E}^{g,k}_{\text{seg}}$. 53
4.5 **Comparison of *ParserNet* with FC-Naive:** We observe that FC-Naive is not able to preserve the fine details and wrinkles present in the single-layered input mesh. Our ParserNet formulation, which is based on a sparse weighted sum of input vertices, preserves the details and is also able to better explain loose clothing than FC-Naive. 57
4.6 ***3D Clothing Parsing* for diverse clothing from input single-layered mesh**. Using ParserNet, one can parse a multi-layered representation of clothing and the body underneath. We present results of ParserNet across different clothing classes and demonstrate that it generalizes well to various clothing styles while preserving fine details. 59
4.7 **ParsetNet+SizerNet module for extracting and resizing clothing**(*Left to Right*) Input single mesh, parsed multi-layer mesh from ParserNet, Resized clothing in two subsequent smaller sizes, heatmap of change in per-vertex error on original parsed clothing for two new sizes. 61
4.8 **ParsetNet+SizerNet module for extracting and resizing clothing**(*Left to Right*): Input single mesh, parsed multi-layer mesh from ParserNet, Resized clothing in two new sizes, heatmap of change in per-vertex error on original parsed clothing for two new sizes. 63
4.9 **Results of *ParserNet* + *SizerNet* for T-Shirt and Shorts**, where we parse the clothing from a input single mesh and change the size of the clothing to visualise dressing effect. 64
4.10 **Results of *ParserNet* + *SizerNet* for Shirt and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect. 64
4.11 **Results of *ParserNet* + *SizerNet* for Coat and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect. 65
4.12 **Results of *ParserNet* + *SizerNet* for Hoodies and Pants**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect. 65
4.13 **Results of *ParserNet* + *SizerNet* for Vest and Shorts**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect. 66
4.14 **Results of *ParserNet* + *SizerNet* for TShirt and Shorts**, where we parse the clothing from input single mesh and change the size of clothing to visualise dressing effect. 66

4.15 **Draping clothes on different body shape:** Given an input clothing in a fixed size, SizerNet can also drape the clothing to different body shapes to see the effect of fixed clothing size on different body shapes. 67
4.16 **Draping clothes on different body shapes and modifying the clothing size:** Given an input clothing in a fixed size, SizerNet can also drape the clothing to different body shapes and further modify the clothing size. . . . 68
4.17 **Interpenetration loss in ParserNet:** Results of ParserNet trained without and with $\mathcal{L}_{\text{interp}}$. We observe that with interpenetrtation loss, the parsed clothing has less intersection with the underlying body. 69
4.18 **Laplacian loss in ParserNet:** Results of ParserNet trained without and with $\mathcal{L}_{\text{lap}}$. We observe that with laplacian loss, the parsed clothing has more realsitic and smooth deformation. 69

5.1 We present **Neural Generalized Implicit Functions** (Neural-GIF), to animate people in clothing as a function of body pose. Neural-GIF learns directly from scans, models complex clothing and produces pose-dependent details for realistic animation. We show for four different characters with the query input pose on the left (illustrated with a skeleton) and our output animation on the right. 73
5.2 **Generalized implicit function evaluation under different transformations:** Here, we use occupancy as an implicit function. (*Left*) Rigid transformation applies rotation and translation$(\mathbf{R}, \mathbf{t})$ mapping canonical points $\mathbf{p}$ to transformed points $\mathbf{p}'$. Occupancy is evaluated by inverse-transforming points back to the canonical space. (*Right*) Displacement field-based transformation, where a spatially varying displacement field $\mathbf{d}(\mathbf{p})$ maps canonical points to displaced points $\mathbf{p}'$. Occupancy is similarly determined by evaluating the original implicit function after inverse displacement. 78
5.3 **Neural Generalized Implicit Function:** Neural-GIF first maps every point $\mathbf{p}$ in 3D posed space to the canonical space by learning the associations between the point and the human body with $f_{\mathbf{w}}$ based on a pose encoding $(\mathbf{P}_\theta)$ as input. Then we predict a displacement field model (f_{def}) learns a pose-dependent deformation field in canonical space $(\Delta\bar{\mathbf{p}}_\theta)$, conditioned on weighted pose parameters $(w \circ \theta)$ and pose encoding $(\bar{\mathbf{P}}_\theta)$ in the canonical space. We add the displacement to the unposed point before evaluating the SDF using our canonical SDF network (f_{CSDF}). We also employ a normal prediction network (f_{norm}) to obtain results with fine-level details. Both f_{CSDF} and f_{norm} take weighted pose and pose encoding as input. 80
5.4 **ClothSeq dataset**: We propose a dataset which consists of subjects in very complicated clothing, such as layered and puffer jackets and open loose jackets in diverse poses. 89
5.5 **Animating people in complex clothes:** Results of Neural-GIF for subjects from the ClothSeq dataset. Neural-GIF can model well complex cloth geometries well, like jacket with hoods (*top*), puffer jacket (*middle*) and loose clothes like shrugs (*bottom*). 92

5.6 **Reanimating CAPE subjects:** Reposing a character from CAPE dataset [MYR*20] using our proposed method. We show results on the walking motion sequence (*top*) and on the playing volleyball motion sequence (*bottom*). 93
5.7 **Comparison with TailorNet [PLPM20] on separate clothing items:** Qualitative comparison on skirt (*top*) and shirt (*bottom*) meshes from the TailorNet dataset. TailorNet produces unsmooth surfaces with unnatural wrinkles, while NeuralGIF generates realistic and high-quality deformations. 95
5.8 **Results on DFAUST [BRPMB17a]:** We show our results on three different motion sequences (running, jiggle on toes, shake hips) from DFAUST and compare with ground-truth DFAUST data and visualize results in (prediction, GT) pairs. Neural-GIF is able to retain pose-dependent soft-tissue deformations present in the ground truth scans. 96
5.9 **Comparison with SCANimate:** We compare the results of our method on the CAPE dataset with SCANimate [SYMB21]. Our model preserves more details and does not have posing artifacts. 97
5.10 **Ablation on learned v/s diffused skinning weights:** Canonical SDF+SMPL weights result in wrong association and hence result in wrong occupancy (*left*). Our canonical mapping network solves this problem by learning blend weights from data but produces smooth results (*middle*). Finally, the displacement field network and normal prediction network help in generating small pose-dependent details (*right*). 99
5.11 **Comparison with part-based model (NASA [DLJ*20]):** We compare the results of our method on the ClothSeq dataset with NASA [DLJ*20]. Our model reconstructs complex clothing without any part-based artifacts and generates realistic pose-dependent deformations. 100
5.12 **Comparison with mesh based method:** Comparison of the raw scan (*left*), SMPL+D registration (*middle*), and our NeuralGIF result (*right*). The registration fails to capture the complex geometry of the open jacket and fine details due to mesh resolution and body-conforming nature of SMPL+D, which also limits the expressivity of models trained on it. 101

6.1 **TailorDiff- A unified model of clothing:** We present TailorDiff, a hierarchical latent point diffusion model for unified and expressive generative modeling of 3D clothing. TailorDiff is topologically flexible and generates diverse clothing for diverse body shapes with a single model (*left*). The hierarchical and body-aware latent space of TailorDiff provides coarse to fine-grained control over generation (*right*). 105
6.2 **TailorDiff Hierarchical VAE:** TailorDiff is structured as a hierarchical VAE with DDM operating in the latent space. Given a clothing pointcloud $\mathbf{c}$ with corresponding body shape ($\mathbf{x_b}$) clothing occupancy on body($\mathbf{o}$), a VAE model is trained which learns a global latent shape ($\mathbf{z} \in \mathbb{R}^{D_z}$) and a local point latent ($\mathbf{h} \in \mathbb{R}^{N \times D_h}$). 110

6.3 **TailorDiff Latent Diffusion Model**: A DDM is trained on the learned clothing latent space. TailorDiff results in diverse and coarse-to-fine-grained control over clothing generation using the learned latent and input conditions. 112
6.4 **Cloth style editing by interpolation for Dress:** Given two input clothing, we show editing by interpolating between the latents. 121
6.5 **Cloth style editing by interpolation for Skirts:** Given two input clothing, we show editing by interpolating between the latents. 122
6.6 **Cloth style editing by interpolation for Jumpsuits:** Given two input clothing, we show editing by interpolating between the latents. 122
6.7 **Comparison with prior work for cloth style editing:** For a fixed body shape, we edit the clothing style by interpolating in latent space. Specifically, the middle column is generated by blending the latent codes corresponding to the clothing styles shown in the left and right columns. 123
6.8 **Fine-grained editing:** Given an input clothing(*left*), we show fine grain-control on the clothing generated. *Top*: We use diffuse-denoise to modify the clothing style, and *Bottom*: We use **o** to control the sleeve style, while keeping other controls fixed. 124
6.9 **More fine-grained editing examples for diverse clothing styles:** Given an input clothing(*left*), we show fine-grained control on the clothing generated. We only change the length here by adjusting the cloth occupancy on the body, while keeping all other controls fixed. 125
6.10 **Failure cases of prior clothing models:** We observe that prior work such as DrapeNet [DLLG*23] often results in unrealistic draping for new shape of body or clothing. SMPLicit [CPA*21a] often results in missing geomtery and big holes. NeuralABC [HYZ24] results in holes and bumpy artefacts near torse. 125
6.11 **Body Shape Editing:** For a given clothing, we change the underlying body shape. TailorDiff adapts perfectly to very diverse body shapes. 127
6.12 **Body shape editing comparison with SMPLicit [CPA*21a].** As shown in the figure, SMPLicit struggles to generate a well-fitting shirt and exhibits collisions with the underlying body, sometimes resulting in holes in the clothing mesh. In contrast, our model adapts effectively to diverse body shapes, producing high-quality results without collisions. 128
6.13 **Body shape editing comparison with NeuralABC [HYZ24].** As shown in the figure, clothing generated by NeuralABC exhibit collisions with the underlying body, particularly in the chest region. In contrast, our TailorDiff avoids such issues, consistently producing well-fitted clothing for diverse body shapes. 128
6.14 **Body shape editing for more different clothing styles:** Our body-aware generation mechanism is capable of adapting to diverse and challenging body shapes, whether short, tall, slim, or overweight. The generated clothing aligns seamlessly with the underlying body and does not result in collisions. 130

6.15 **Body shape editing for more different clothing styles:** Our body-aware generation mechanism is capable of adapting to diverse and challenging body shapes, whether short, tall, slim, or overweight. The generated clothing aligns seamlessly with the underlying body and does not result in collisions. 131
6.16 **Clothing reconstruction of Skirts and Shirts:** We visualize the target 3D clothing (*left*), the reconstruction by ISP [LGF23] (*middle*), and the reconstruction by our method (*right*). ISP struggles to accurately reconstruct the clothing, resulting in collisions with the body. In contrast, our body-aware mechanism effectively avoids such limitations. 132
6.17 **Layered scan fitting comparison with prior works:** We compare layered scan fitting with GALA [KKSJ24] and DrapeNet [DLLG*23]. DrapeNet struggles to model 3-layer clothing. GALA creates volumetric meshes that do not reflect the reality of the clothing surface, and also result in significant intersections. 134
6.18 **Layered scan fitting on more scans with diverse clothing:** TailorDiff based fitting enables layer-wise fitting across a wide variety of clothing types, something that prior works cannot achieve. 135
6.19 **Layered clothing assets from scans:** Given an input scan with segmentation labels(*top*), where only partial surfaces of the inner clothing are visible, our scan fitting method effectively fits multiple clothing layers to the scan while preserving the style and structure of the clothing. In contrast, GALA [KKSJ24] can achieve similar results but produces noisy meshes with incorrect shapes (*see Extracted layers*) and significant intersections between layers (*see Layered visualization*). Our method, however, generates accurate shapes and styles of clothing with proper layer ordering, ensuring no intersections between layers. 136
6.20 **Impact of body-aware encoding:** Body-aware encoding results in better alignment of clothing and body and also yields less intersection. 137

7.1 **Pose-NDF-a neural unsigned distance field in $SO(3)^K$ space:** Pose-NDFlearns the manifold of plausible poses as zero level set. We learn the distance field representation from samples of plausible (■) and unrealistic (●) poses (*left*). We encode the input pose (given as a set of quaternions) using a structural MLP f^{enc} and predict the distance from the joint representation using an MLP f^{df}. The gradient $\nabla_\theta f^{\text{udf}}$ and distance value $f^{\text{udf}}(\theta)$ are used to project implausible poses onto the manifold (*right*). 144
7.2 **Hierarchical Implicit Neural Function:** $f^{\text{udf}}(\theta) = (f^{\text{df}} \circ f^{\text{enc}})(\theta)$ First we encode the pose θ_k using f_k^{enc}, where $k \in \{2 \dots K\}$. Each joint pose is embedded using an MLP f_k^{enc}, conditioned on the parent features $\mathbf{v}_{\tau(k)}$. The resulting embeddings $\mathbf{v}_k$ are concatenated to form the pose representation $\mathbf{p}$, which is processed by f^{df} to predict unsigned distance fields. For visuslisation, we use a skeleton with less joints. 148

7.3 **Comparison with prior works on motion denoising**: We observe that Pose-NDF based motion denoising makes the pose realistic and solves small intersection issues, while VPoser and HuMoR still result in unrealistic poses. 154
7.4 **Motion denoising using Pose-NDF v/s HuMoR [RBH*21]**: We compare the results on motion denoising using Pose-NDF and HuMoR [RBH*21] as priors with GT data and visualize every 10^{th} frame of a sequence. We observe that for HuMoR (*top*), the correction in input pose accumulates over time and makes the output pose significantly different from the GT (*middle*). Pose-NDF remains close to observations while correcting unrealistic poses (*bottom*). 156
7.5 **Motion denoising for noisy mocap data ("Noissy AMASS"):** We observe that Pose-NDF based motion denoising makes the pose realistic and also resembles the input observation, while VPoser and HuMoR still result in unrealistic poses and, in some cases, deviate too much from the input observation (*e.g. bottom-left*). 158
7.6 **Fitting to partial observation:** We compare VPoser [PCG*19], HuMoR [RBH*21], and Pose-NDF based prior on the task of recovering 3D pose from partial observation. *Top*: In case of occluded left leg, we observe that VPoser produces nearly straight legs(close to mean position), while HuMoR and PoseNDF produce much more diverse poses. HuMoR results in extremely unrealistic poses in some cases. *Bottom*: We observe similar behavior for the occluded arm case. 160
7.7 **Human pose and shape estimation from in-the-wild images:** We use Pose-NDF as a prior in optimization based method to estimate 3D pose and shape from images. 161
7.8 **Diversity and realism in pose generation:** GMM tends to produce highly unrealistic poses, while VPoser, GAN-S, and Pose-NDF generate significantly more realistic ones. According to the APD metric, Pose-NDF exhibits greater pose variance compared to VPoser and GAN-S, leading to a higher diversity in the generated poses. 163
7.9 **Comparison with VPoser [PCG*19] on pose interpolation:** We observe that VPoser-based interpolation (*top*) is less smooth than Pose-NDF-based interpolation (*bottom*). 164
7.10 **Pose Interpolation using Pose-NDF:** We observe smooth and more consistent interpolation using Pose-NDF. 166

List of tables

3.1 **Real-world Human in Clothing Datasets:** We present an overview of 3D(*top*) and 4D(*bottom*) scans of humans in clothing, comparing the datasets in terms of their variations and the types of annotations they provide. 24

3.2 **Synthetic Clothing Datasets:** Synthetic datasets consist of 3D Clothing simulated on different body shapes and poses and consist of 3D static scans (*top*) or dynamic 4D sequences (*middle*). Some recent datasets also provide loose clothing in their dataset. **TailorDiff**(*bottom*) combines diverse datasets by aligning meshes to a common frame and registering them to the SMPL model. 26

4.1 **Comparison of SIZER with existing datasets:** We compare with 3D and 4D human/clothing datasets published prior to its release. SIZER dataset uniquely combines a wide range of clothing types with size variation across a large number of subjects, enabling size-sensitive clothing modeling. . . . 49

4.2 **Clothing size variation using per-vertex differences:** Average per-vertex displacement (in mm) between clothing items of different sizes. The differences are nonlinear and vary with both the source size and the type of clothing. Transitions from smaller sizes to larger ones, especially for loose clothing like shorts and hoodies, show larger vertex displacements than transitions between adjacent larger sizes. 51

4.3 **Clothing size variation using surface area:** Percentage change in clothing surface area across different size transitions. The change is nonlinear and varies significantly depending on both the source and target sizes as well as the clothing type. 52

4.4 **Average per-vertex error V_{err} for *3D Clothing Parsing*:** We compare the parsed clothing with ground-truth clothing meshes registered to a predefined common template across each clothing class. We observe that ParserNet achieves better results than Linear Model and FC-Naive. All the numbers are reported in *mm*. 58

4.5 **Average per-vertex error (V_{err} in mm) and surface area error (A_{err} in %) for clothing resizing:** We compare the aforementioned metrics between clothing predicted using SizerNet and the ground-truth clothing meshes, which are registered to a common template for each clothing class. We evaluate our model against the ground-truth error margin across all sizes, the average prediction, and a linear scaling model. These results suggest that clothing resizing is a highly complex task and requires data-driven deep models to learn the underlying deformations. 60

4.6 **ParserNet Ablation:** Impact of surface regulariser and intersection loss in ParserNet predictions. 68

5.1 **Comparison of skinning approaches used in *human in clothing* models:** While mesh-based LBS offers simple and low computational cost, implicit methods can model complex and highly dynamic deformation, with the backward skinning-based formulation excelling in capturing dynamics and clothing deformation. 86
5.2 **Pose blend shapes and neural displacement fields for modeling pose-dependent deformations:** While SMPL+D uses explicit vertex-based blend shapes with low computational cost, NeuralGIF leverages an implicit volumetric field, offering higher topological flexibility and fine-detail modeling. 87
5.3 **SMPL+D and Neural-GIF models for clothed human representation:** SMPL+D uses explicit mesh-based deformations with limited flexibility, while Neural-GIF employs implicit volumetric fields, enabling flexible topology, high-detail modeling, and direct training from raw scans, albeit at a higher computational cost. 87
5.4 **Comparison of ClothSeq with other 4D human scan datasets:** Our dataset offers real scans of complex clothing across multiple subjects and temporal sequences, presenting a more realistic and challenging benchmark for dynamic clothing modeling. Datasets released prior to ClothSeq are shaded in grey. Some new datasets like 4D-DRESS [WHG*24], Actors-HQ [IRG*23a] have very loose clothing. 88
5.5 **Comparison of NeuralGIF with prior and concurrent works:** We quantitatively compare the results of our method with NASA [DLJ*20] and SCANimate [SYMB21]. We report point to surface distance (in mm) and IoU and F-Scores(%) for comparison. 95
5.6 **Ablation on learned v/s diffused skinning weights:** Quantitative comparison between using nearest-neighbor-based SMPL weights and our learned canonical mapping network for predicting skinning weights. Our learned skinning weights consistently outperform the baseline, significantly improving reconstruction accuracy and surface alignment. 98
5.7 **Comparison with multi-shape model (LEAP [MZBT21]):** Quantitative results on three different datasets(DFAUST, MoVi, and SMPL).Neural-GIF achieves competitive or superior performance, demonstrating its effectiveness in capturing pose-dependent surface deformations. 100

6.1 **TailorDiff v/s prior works for unified model of clothing:** Comparison of our method with other works. Ours is the first unified model for generating clothing of all topologies and categories using a single model. 108

6.2 **Dataset Overview:** Details of publicly available datasets used for training our model. We combine all these datasets by aligning them to a single body model (SMPL in our case) and posing all the clothing scans in a fixed canonical pose. Finally, the samples mentioned in the last column are high-quality, well-aligned scans that are used to train and evaluate our model. 114
6.3 **Generation metric:** Comparison of all generation metrics, JSD, COV, MMD 1-NNA across four categories: Upper Clothing, Lower Clothing, Skirts and Dress/Jumpsuits, for different models. (*Note: Empty entries in the table indicate cases that are not feasible to execute with the previous methods.*) . 117
6.4 **Body-Cloth Intersection:** We compare the body-cloth intersection percentage (% ↓) across different models and clothing categories. 119
6.5 **Clothing reconstruction:** Evaluation results of clothing reconstruction on CLOTH3D test set (*in cm*). 133
6.6 **Ablation on network architecture and latent features:** We evaluate the impact of body encoding and a hierarchical latent space on diversity (1-NNA %) and realism (body-cloth intersection). 137

7.1 **Motion denoising**: We compare the per-vertex error (in *cm*) on mocap data from HPS (*left*) and AMASS (*middle*) and on artificially created noisy AMASS data (*right*). In all cases, Pose-NDF based motion denoising results in the least error. We also observe that in the case of mocap data (HPS, AMASS), motion denoising using Pose-NDF results in a very small error (a small change from the input), which is the desired behavior, as these mocap poses are already realistic and hence close to our learned manifold. On the other hand, HuMoR significantly changes the input pose. 153
7.2 **Pose realism using self-intersection:** Average % of self-intersecting mesh faces in motion denoising. It can be seen that in most cases, Pose-NDF produces fewer intersections than VPoser and HuMoR. 157
7.3 **Motion estimation from partial 3D observations**: We compare per-vertex error (in *cm*). It can be seen that for leg and arm/hand occlusions, Pose-NDF reconstructs the pose better than VPoser and HuMoR. For occluded shoulders, HuMoR takes the lead. We observe that the results of Pose-NDF depend on the initialization of the occluded joint, as it is expected from manifold projection. 159
7.4 **Optimization-based 3D pose and shape estimation.** Comparison of optimization methods using different priors: VPoser [PCG*19], GAN-S [DRC*22], and Pose-NDF. 162
7.5 **ExPose and ExPose + Optimization results.** We evaluate the proposed prior and optimization pipeline to enhance the SoTA 3D pose and shape estimation network, ExPose [CPB*20]. 162
7.6 **Diversity and realism of generated poses from different priors:** We report Average Pairwise Distance (APD) in cm and percentage of self-intersecting mesh faces for 500 randomly sampled poses. Higher APD indicates more diversity, while a lower intersection percentage indicates better realism. . . 164

7.7 **Smoothness in pose interpolation:** Mean per-vertex distances (± standard deviation) between consecutive poses. Lower values indicate smoother interpolations. 165
7.8 **Cumulative error across deviation levels from the mean pose (σ, 2σ, 3σ):** We evaluate per-vertex reconstruction error(in cm) for AMASS Noisy data (60 and 120 frames). We compare the performance of Pose-NDF, VPoser, and HuMoR on motion denoising. 165

References

[3DL25] 3DLOOK: 3dlook – ai-driven body scanning and virtual fitting technology, 2025. Accessed: 2025-05-06.

[3dM] 3dMD.

[AB15] AKHTER I., BLACK M. J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR* (2015).

[ABB*23] ARMANDO M., BOISSIEUX L., BOYER E., FRANCO J.-S., HUMENBERGER M., LEGRAS C., LEROY V., MARSOT M., PANSIOT J., PUJADES S., REKIK R., ROGEZ G., SWAMY A., WUHRER S.: 4dhumanoutfit: a multi-subject 4d dataset of human motion sequences in varying outfits exhibiting large displacements. *Computer Vision and Image Understanding* (2023).

[ACP03] ALLEN B., CURLESS B., POPOVIĆ Z.: The space of human body shapes: reconstruction and parameterization from range scans. *ACM Trans. Graph. 22*, 3 (July 2003), 587–594.

[ACPH06] ALLEN B., CURLESS B., POPOVIC Z., HERTZMANN A.: Learning a Correlated Model of Identity and Pose-Dependent Body Shape Variation for Real-Time Synthesis. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation* (2006), Cani M.-P., O'Brien J., (Eds.), The Eurographics Association.

[ADMG18] ACHLIOPTAS P., DIAMANTI O., MITLIAGKAS I., GUIBAS L.: Learning representations and generative models for 3d point clouds, 2018.

[Adv] ADVANCED COMPUTING CENTER FOR THE ARTS AND DESIGN: ACCAD MoCap Dataset. URL: https://accad.osu.edu/research/motion-lab/mocap-system-and-data.

[agi] Agisoft metashape. URL: https://www.agisoft.com/.

[AKH19] AKSAN E., KAUFMANN M., HILLIGES O.: Structured prediction helps 3D human motion modelling. In *ICCV* (2019).

[ALAC21] ANDREOU N., LAZAROU A., ARISTIDOU A., CHRYSANTHOU Y.: A hierarchy-aware pose representation for deep character animation, 2021.

[AMB*19] ALLDIECK T., MAGNOR M., BHATNAGAR B. L., THEOBALT C., PONS-MOLL G.: Learning to reconstruct people in clothing from a single RGB camera. In *CVPR* (2019).

[AMX*18a] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Detailed human avatars from monocular video. In *International Conference on 3D Vision (3DV)* (Sept. 2018).

[AMX*18b] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Video based reconstruction of 3D people models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018).

[APMTM19] ALLDIECK T., PONS-MOLL G., THEOBALT C., MAGNOR M.: Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)* (Oct. 2019), IEEE.

[ASK*05] ANGUELOV D., SRINIVASAN P., KOLLER D., THRUN S., RODGERS J., DAVIS J.: Scape: Shape completion and animation of people. *ACM Trans. Graph. 24*, 3 (July 2005), 408–416.

[ASSS*20] ALIAKBARIAN S., SADAT SALEH F., SALZMANN M., PETERSSON L., GOULD S.: A stochastic conditioning scheme for diverse human motion prediction. In *CVPR* (2020).

[ATO*24] ANTIĆ D., TIWARI G., OZCOMLEKCI B., MARIN R., PONS-MOLL G.: CloSe: A 3D clothing segmentation dataset and model. In *International Conference on 3D Vision (3DV)* (Mar. 2024).

[AWH*22] AGGARWAL A., WANG J., HOGUE S., NI S., BUDAGAVI M., GUO X.: Layered-garment net: Generating multiple implicit garment layers from a single image, 2022.

[axy] AXYZ design. URL: https://secure.axyz-design.com/.

[BB00] BAERLOCHER P., BOULIC R.: Parametrization and range of motion of the ball-and-socket joint. In *Proceedings of the IFIP TC5/WG5.10 DEFORM'2000 Workshop and AVATARS'2000 Workshop on Deformable Avatars* (2000).

[BB08] BĂLAN A. O., BLACK M. J.: The naked truth: Estimating body shape under clothing. In *European Conf. on Computer Vision* (2008), Springer, pp. 15–29.

[BBR*22] BAI Z., BAGAUTDINOV T., ROMERO J., ZOLLHÖFER M., TAN P., SAITO S.: Autoavatar: Autoregressive neural fields for dynamic avatar modeling. In *European conference on computer vision* (2022).

[BKL*16] BOGO F., KANAZAWA A., LASSNER C., GEHLER P., ROMERO J., BLACK M. J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV* (2016).

[BKL18] BARSOUM E., KENDER J., LIU Z.: HP-GAN: probabilistic 3D human motion prediction via gan. In *CVPR Workshops* (2018).

[BME19] BERTICHE H., MADADI M., ESCALERA S.: CLOTH3D: clothed 3D humans. vol. abs/1912.02792.

[BME20] BERTICHE H., MADADI M., ESCALERA S.: Cloth3d: clothed 3d humans. In *European Conference on Computer Vision* (2020), Springer, pp. 344–359.

[BME22] BERTICHE H., MADADI M., ESCALERA S.: Neural cloth simulation. *ACM Trans. Graph. 41*, 6 (nov 2022).

[BMTE21] BERTICHE H., MADADI M., TYLSON E., ESCALERA S.: Deepsd: Automatic deep skinning and pose space deformation for 3d garment animation, 2021.

[BP07] BARAN I., POPOVIĆ J.: Automatic rigging and animation of 3d characters. *ACM Trans. Graph. 26*, 3 (July 2007), 72–es.

[BPS*08] BRADLEY D., POPA T., SHEFFER A., HEIDRICH W., BOUBEKEUR T.: Markerless garment capture. In *ACM Transactions on Graphics* (2008), vol. 27, ACM, p. 99.

[BRLB14] BOGO F., ROMERO J., LOPER M., BLACK M. J.: FAUST: Dataset and evaluation for 3D mesh registration. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Columbus, Ohio, USA, June 2014), pp. 3794 –3801. doi:10.1109/CVPR.2014.491.

[BRPMB17a] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition* (2017).

[BRPMB17b] BOGO F., ROMERO J., PONS-MOLL G., BLACK M. J.: Dynamic FAUST: Registering human bodies in motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[BSTPM20a] BHATNAGAR B. L., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)* (Aug. 2020), Springer.

[BSTPM20b] BHATNAGAR B. L., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Advances in Neural Information Processing Systems (NeurIPS)* (Dec. 2020).

[BTH*03] BHAT K. S., TWIGG C. D., HODGINS J. K., KHOSLA P. K., POPOVIĆ Z., SEITZ S. M.: Estimating cloth simulation parameters from video. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Goslar, DEU, 2003), SCA '03, Eurographics Association, p. 37–51.

[BTTPM19] BHATNAGAR B. L., TIWARI G., THEOBALT C., PONS-MOLL G.: Multi-garment net: Learning to dress 3D people from images. In *ICCV* (2019).

[BW98] BARAFF D., WITKIN A.: Large steps in cloth simulation. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1998), SIGGRAPH '98, Association for Computing Machinery, p. 43–54.

[BXP*22] BHATNAGAR B. L., XIE X., PETROV I. A., SMINCHISESCU C., THEOBALT C., PONS-MOLL G.: Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022).

[BXX*25] BIAN S., XU C., XIU Y., GRIGOREV A., LIU Z., LU C., BLACK M. J., FENG Y.: Chatgarment: Garment estimation, generation and editing via large language models.

[CAPM20] CHIBANE J., ALLDIECK T., PONS-MOLL G.: Implicit functions in feature space for 3D shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020), IEEE.

[CAPMMN24] CORONA E., ALENYA G., PONS-MOLL G., MORENO-NOGUER F.: LayerNet: High-Resolution Semantic 3D Reconstruction of Clothed People . *IEEE Transactions on Pattern Analysis & Machine Intelligence 46*, 02 (Feb. 2024).

[Car] CARNEGIE MELLON UNIVERSITY: CMU MoCap Dataset. URL: http://mocap.cs.cmu.edu.

[CCH*23] CAO Y., CAO Y.-P., HAN K., SHAN Y., WONG K.-Y. K.: Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. *arXiv preprint arXiv:2304.00916* (2023).

[CCSC23] CHEN C.-Y., CHEN Y.-C., SHUAI H.-H., CHENG W.-H.: Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2023), pp. 7513–7522.

[CHS*19] CAO Z., HIDALGO MARTINEZ G., SIMON T., WEI S., SHEIKH Y. A.: OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[CJS*22] CHEN X., JIANG T., SONG J., YANG J., BLACK M. J., GEIGER A., HILLIGES O.: gdna: Towards generative detailed neural avatars. *arXiv* (2022).

[CJS*23] CHEN X., JIANG T., SONG J., RIETMANN M., GEIGER A., BLACK M. J., HILLIGES O.: Fast-snarf: A fast deformer for articulated neural fields. *Pattern Analysis and Machine Intelligence (PAMI)* (2023).

[CK02] CHOI K.-J., KO H.-S.: Stable but responsive cloth. *ACM Transactions on Graphics 21* (07 2002).

[CL83] CRANDALL M. G., LIONS P.-L.: Viscosity solutions of Hamilton-Jacobi equations. *Transactions of the American mathematical society 277*, 1 (1983), 1–42.

[CL24] CHEN R. T. Q., LIPMAN Y.: Flow matching on general geometries, 2024.

[CLI*20] CHABRA R., LENSSEN J. E., ILG E., SCHMIDT T., STRAUB J., LOVEGROVE S., NEWCOMBE R. A.: Deep local shapes: Learning local SDF priors for detailed 3D reconstruction. In *ECCV* (2020).

[CLLY22] CHEN W., LIN C., LI W., YANG B.: 3psdf: Three-pole signed distance function for learning surfaces with arbitrary topologies. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (June 2022).

[CMPM20] CHIBANE J., MIR A., PONS-MOLL G.: Neural unsigned distance fields for implicit function learning. In *Advances in Neural Information Processing Systems (NeurIPS)* (Dec. 2020).

[CPA*21a] CORONA E., PUMAROLA A., ALENYÀ G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *CVPR* (2021).

[CPA*21b] CORONA E., PUMAROLA A., ALENYA G., PONS-MOLL G., MORENO-NOGUER F.: Smplicit: Topology-aware generative model for clothed people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), IEEE.

[CPB*20] CHOUTAS V., PAVLAKOS G., BOLKART T., TZIONAS D., BLACK M. J.: Monocular expressive body regression through body-driven attention. In *ECCV* (2020).

[CPM20] CHIBANE J., PONS-MOLL G.: Neural distance fields for unconstrained shape representation. *Advances in Neural Information Processing Systems (NeurIPS)* (2020).

[CPMAMN22] CORONA E., PONS-MOLL G., ALENYA G., MORENO-NOGUER F.: Learned vertex descent: A new direction for 3D human model fitting. In *European Conference on Computer Vision (ECCV)* (Oct. 2022), Springer.

[CPW*21] CHEN X., PANG A., WEI Y., PEIHAO W., XU L., YU J.: Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (Presented at ACM SIGGRAPH)* (2021).

[CPZ*19] CHEN X., PANG A., ZHU Y., LI Y., LUO X., ZHANG G., WANG P., ZHANG Y., LI S., YU J.: Towards 3D human shape recovery under clothing. *CoRR abs/1904.02601* (2019).

[CYAE*20] CAI R., YANG G., AVERBUCH-ELOR H., HAO Z., BELONGIE S., SNAVELY N., HARIHARAN B.: Learning gradient fields for shape generation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2020).

[CZB*21] CHEN X., ZHENG Y., BLACK M. J., HILLIGES O., GEIGER A.: SNARF: differentiable forward skinning for animating non-rigid neural implicit shapes. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021* (2021), IEEE, pp. 11574–11584.

[dAST*08] DE AGUIAR E., STOLL C., THEOBALT C., AHMED N., SEIDEL H., THRUN S.: Performance capture from sparse multi-view video. *ACM Trans. Graph. 27*, 3 (2008), 98:1–98:10.

[DCKY02] DONG F., CLAPWORTHY G. J., KROKOS M. A., YAO J.: An anatomy-based approach to human muscle modeling and deformation. *IEEE Transactions on Visualization and Computer Graphics 8*, 2 (2002).

[DFH*24] DONG J., FANG Q., HUANG Z., XU X., WANG J., PENG S., DAI B.: Tela: Text to layer-wise 3d clothed human generation, 2024.

[DFZ*24] DANG Z., FAN T., ZHAO B., SHEN X., WANG L., ZHANG G., CUI Z.: Momanifold: Learning to measure 3d human motion via decoupled joint acceleration manifolds. In *Proceedings of the British Machine Vision Conference* (09 2024).

[DLJ*20] DENG B., LEWIS J., JERUZALSKI T., PONS-MOLL G., HINTON G., NOROUZI M., TAGLIASACCHI A.: Nasa: Neural articulated shape approximation. In *The European Conference on Computer Vision (ECCV)* (Aug. 2020).

[DLLG*23] DE LUIGI L., LI R., GUILLARD B., SALZMANN M., FUA P.: DrapeNet: Garment Generation and Self-Supervised Draping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023).

[DLW*19] DONG H., LIANG X., WANG B., LAI H., ZHU J., YIN J.: Towards multi-pose guided virtual try-on network. *International Conference on Computer Vision (ICCV)* (2019).

[DLZ*20] DONG H., LIANG X., ZHANG Y., ZHANG X., XIE Z., WU B., ZHANG Z., SHEN X., YIN J.: Fashion editing with adversarial parsing learning. *Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).

[DMQ*23] DAI L., MA L., QIAN S., LIU H., LIU Z., XIONG H.: Cloth2body: Generating 3d human body mesh from 2d clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 15007–15017.

[DRC*22] DAVYDOV A., REMIZOVA A., CONSTANTIN V., HONARI S., SALZMANN M., FUA P.: Adversarial parametric pose prior. In *CVPR* (2022).

[ENNE12] ENGELL-NØRREGÅRD M., NIEBE S., ERLEBEN K.: A joint-constraint model for human joints using signed distance-fields. *Multibody System Dynamics 28* (2012).

[EST*20] EGGER B., SMITH W. A. P., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., THEOBALT C., BLANZ V., VETTER T.: 3d morphable face models—past, present, and future. *ACM Trans. Graph. 39*, 5 (June 2020).

[FKL*23] FENG H., KULITS P., LIU S., BLACK M. J., ABREVAYA V.: Generalizing neural human fitting to unseen poses with articulated se(3) equivariance, 2023.

[FLB*23] FENG Y., LIU W., BOLKART T., YANG J., POLLEFEYS M., BLACK M. J.: Learning disentangled avatars with hybrid 3d representations. *arXiv* (2023).

[FYP*22] FENG Y., YANG J., POLLEFEYS M., BLACK M. J., BOLKART T.: Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers* (2022), SA '22.

[GCS*19] GUNDOGDU E., CONSTANTIN V., SEIFODDINI A., DANG M., SALZMANN M., FUA P.: Garnet: A two-stream network for fast and accurate 3D cloth draping. In *ICCV* (2019).

[GLK*20] GEORGAKIS G., LI R., KARANAM S., CHEN T., KOŠECKÁ J., WU Z.: Hierarchical kinematic human mesh recovery. In *ECCV* (2020).

[GLL*18] GONG K., LIANG X., LI Y., CHEN Y., YANG M., LIN L.: Instance-level human parsing via part grouping network. In *ECCV* (2018).

[GM87] GEMAN S., MCCLURE D. E.:. In *Statistical methods for tomographic image reconstruction* (1987).

[GMSPM21] GUZOV V., MIR A., SATTLER T., PONS-MOLL G.: Human POSEitioning System (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR* (2021).

[GMT*20] GHORBANI S., MAHDAVIANI K., THALER A., KORDING K., COOK D. J., BLOHM G., TROJE N. F.: MoVi: A large multipurpose motion and video dataset. *arXiv preprint arXiv: 2003.01888* (2020).

[GMT*21] GHORBANI S., MAHDAVIANI K., THALER A., KORDING K., COOK D. J., BLOHM G., TROJE N. F.: Movi: A large multi-purpose human motion and video dataset. *PLOS ONE 16* (June 2021), e0253157.

[GNF*25] GAN Q., NGUYEN S. M., FENAUX E., CLÉMENÇON S., YACOUBI M. E.: Polar coordinate-based 2d pose prior with neural distance field, 2025.

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *NeurIPS* (2014).

[GRH*12] GUAN P., REISS L., HIRSHBERG D., WEISS A., BLACK M. J.: DRAPE: DRessing Any PErson. *ACM Trans. on Graphics (Proc. SIGGRAPH) 31*, 4 (July 2012), 35:1–35:10.

[GSF22] GUILLARD B., STELLA F., FUA P.: Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *European Conference on Computer Vision* (2022).

[GTBH23] GRIGOREV A., THOMASZEWSKI B., BLACK M. J., HILLIGES O.: HOOD: Hierarchical graphs for generalized modelling of clothing dynamics.

[GYH*20] GROPP A., YARIV L., HAIM N., ATZMON M., LIPMAN Y.: Implicit geometric regularization for learning shapes. In *Proceedings of Machine Learning and Systems* (2020).

[GZW*20] GUO C., ZUO X., WANG S., ZOU S., SUN Q., DENG A., GONG M., CHENG L.: Action2Motion: Conditioned generation of 3d human motions. In *ACMMM* (2020).

[GZZ*22] GUO C., ZOU S., ZUO X., WANG S., JI W., LI X., CHENG L.: Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 5152–5161.

[HBVT99] HADAP S., BANGERTER E., VOLINO P., THALMANN N.: Animating wrinkles on clothes. pp. 175–182.

[HCJS20] HE T., COLLOMOSSE J. P., JIN H., SOATTO S.: Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *ArXiv* (2020).

[HGZ*23] HU L., GAO X., ZHANG P., SUN K., ZHANG B., BO L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117* (2023).

[HHH*23] HU S., HONG F., HU T., PAN L., MEI H., XIAO W., YANG L., LIU Z.: Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint* (2023).

[HILJO23] HSUAN-I H., LIXIN X., JIE S., OTMAR H.: Learning locally editable virtual humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2023).

[HL23] HU S., LIU Z.: Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:* (2023).

[HLRB12] HIRSHBERG D. A., LOPER M., RACHLIN E., BLACK M. J.: Coregistration: simultaneous alignment and modeling of articulated 3d shape. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI* (Berlin, Heidelberg, 2012), ECCV'12, Springer-Verlag, p. 242–255.

[HPY*23] HAN S.-H., PARK M.-G., YOON J. H., KANG J.-M., PARK Y.-J., JEON H.-G.: High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)* (June 2023).

[HSS*09] HASLER N., STOLL C., SUNKEL M., ROSENHAHN B., SEIDEL H.-P.: A statistical model of human pose and body shape. *Computer Graphics Forum 28*, 2 (2009), 337–346.

[HTB*24] HE Y., TIWARI G., BIRDAL T., LENSSEN J. E., PONS-MOLL G.: Nrdf: Neural riemannian distance fields for learning articulated pose priors. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024).

[HTRM15] HELM F., TROJE N., REISER M., MUNZERT J.: Bewegungsanalyse getÃ¤uschter und nicht-getÃ¤uschter 7m-wÃ¼rfe im handball.

[HTZ*25] HE Y., TIWARI G., ZHANG X., BORA P., BIRDAL T., LENSSEN J. E., PONS-MOLL G.: Limo: Lifelike human motion generation with continuous-space generative models. Under Review.

[HUF04] HERDA L., URTASUN R., FUA P.: Hierarchical implicit surface joint limits for human body tracking. In *ECCV* (2004).

[HUH02] HERDA L., URTASUN R., HANSON A.: Automatic determination of shoulder joint limits using quaternion field boundaries. In *FG* (2002).

[Huy09] HUYNH D. Q.: Metrics for 3D rotations: Comparison and analysis. *J. Math. Imaging Vis. 35*, 2 (Oct. 2009), 155–164.

[HX*19] HABERMANN M., XU W., , ZOLLHOEFER M., PONS-MOLL G., THEOBALT C.: Livecap: Real-time human performance capture from monocular video.

[HX*20] HABERMANN M., XU W., , ZOLLHOEFER M., PONS-MOLL G., THEOBALT C.: Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020), IEEE.

[HXL*20] HUANG Z., XU Y., LASSNER C., LI H., TUNG T.: Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 3093–3102.

[HXS*21] HE T., XU Y., SAITO S., SOATTO S., TUNG T.: Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021).

[HYH*20] HEMING Z., YU C., HANG J., WEIKAI C., DONG D., ZHANGYE W., SHUGUANG C., XIAOGUANG H.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision – ECCV 2020* (2020), Springer International Publishing.

[HYL*21] HE Y., YU H., LIU X., YANG Z., SUN W., WANG Y., FU Q., ZOU Y., MIAN A.: Deep learning based 3d segmentation: A survey. *CoRR abs/2103.05423* (2021).

[HYZ24] HONGHU C., YUXIN Y., ZHANG J.: Neural-abc: Neural parametric models for articulated body with clothes. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[HZJ*21] HONG Y., ZHANG J., JIANG B., GUO Y., LIU L., BAO H.: Stereopifu: Depth aware clothed human digitization via stereo vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021).

[HZZ*24] HU L., ZHANG H., ZHANG Y., ZHOU B., LIU B., ZHANG S., NIE L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[ihl] Ihl group. URL: https://www.ihlservices.com/.

[IRG*23a] IŞIK M., RÜNZ M., GEORGOPOULOS M., KHAKHULIN T., STARCK J., AGAPITO L., NIESSNER M.: Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–12.

[IRG*23b] IŞIK M., RÜNZ M., GEORGOPOULOS M., KHAKHULIN T., STARCK J., AGAPITO L., NIESSNER M.: Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–12.

[JBS*23] JUNG H., BRASCH N., SONG J., PEREZ-PELLITERO E., ZHOU Y., LI Z., NAVAB N., BUSAM B.: Deformable 3d gaussian splatting for animatable human avatars, 2023.

[JDJ19] JOHNSON J., DOUZE M., JÉGOU H.: Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data 7*, 3 (2019), 535–547.

[JDKL14] JACOBSON A., DENG Z., KAVAN L., LEWIS J.: Skinning: Real-time shape deformation. In *ACM SIGGRAPH 2014 Courses* (2014).

[JE10] JOHNSON S., EVERINGHAM M.: Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC* (2010).

[JKS13] JACOBSON A., KAVAN L., SORKINE O.: Robust inside-outside segmentation using generalized winding numbers. *ACM Trans. Graph. 32*, 4 (2013).

[JLJ*20] JERUZALSKI T., LEVIN D. I. W., JACOBSON A., LALONDE P., NOROUZI M., TAGLIASACCHI A.: Nilbs: Neural inverse linear blend skinning, 2020.

[JSP*23] JINKA S. S., SRIVASTAVA A., POKHARIYA C., SHARMA A., NARAYANAN P. J.: SHARP: shape-aware reconstruction of people in loose clothing. *Int. J. Comput. Vis. 131*, 4 (2023), 918–937.

[JXY*18] JI Y., XU F., YANG Y., SHEN F., SHEN H. T., ZHENG W.-S.: A large-scale RGB-D database for arbitrary-view human action recognition. In *ACM International Conference on Multimedia (ACMMM)* (2018).

[JYS*22] JIANG W., YI K. M., SAMEI G., TUZEL O., RANJAN A.: Neuman: Neural human radiance field from a single video, 2022.

[JZH*20] JIANG B., ZHANG J., HONG Y., LUO J., LIU L., BAO H.: Bcnet: Learning body and cloth shape from a single image. *arXiv preprint arXiv:2004.00214* (2020).

[JZL*24] JIANG N., ZHANG Z., LI H., MA X., WANG Z., CHEN Y., LIU T., ZHU Y., HUANG S.: Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1737–1747.

[KAB20] KOCABAS M., ATHANASIOU N., BLACK M. J.: Vibe: Video inference for human body pose and shape estimation. In *CVPR* (2020).

[KAC*24] KOLOTOUROS N., ALLDIECK T., CORONA E., BAZAVAN E. G., SMINCHISESCU C.: Instant 3d human avatar generation using image diffusion models. In *European Conference on Computer Vision (ECCV)* (2024).

[KBH06] KAZHDAN M., BOLITHO M., HOPPE H.: Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing* (2006), pp. 61–70.

[KBJM18] KANAZAWA A., BLACK M. J., JACOBS D. W., MALIK J.: End-to-end recovery of human shape and pose. In *CVPR* (2018).

[KBM*24] KHIRODKAR R., BAGAUTDINOV T., MARTINEZ J., ZHAOEN S., JAMES A., SELEDNIK P., ANDERSON S., SAITO S.: Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569* (2024).

[KHTG24] KAIRANDA N., HABERMANN M., THEOBALT C., GOLYANIK V.: Neuralclothsim: Neural deformation fields meet the thin shell theory. In *Neural Information Processing Systems (NeurIPS)* (2024).

[KJP02] KRY P. G., JAMES D. L., PAI D. K.: Eigenskin: real time large deformation character skinning in hardware. In *Proceedings of the 2002 ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (New York, NY, USA, 2002), SCA '02, Association for Computing Machinery, p. 153–159.

[KKK*24] KOROSTELEVA M., KESDOGAN T. L., KEMPER F., WENNINGER S., KOLLER J., ZHANG Y., BOTSCH M., SORKINE-HORNUNG O.: GarmentCodeData: A dataset of 3D made-to-measure garments with sewing patterns. In *Computer Vision – ECCV 2024* (2024).

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics 42*, 4 (July 2023).

[KKSJ24] KIM T., KIM B., SAITO S., JOO H.: Gala: Generating animatable layered assets from a single scan, 2024.

[KL21] KOROSTELEVA M., LEE S.-H.: Generating datasets of 3d garments with sewing patterns. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (2021).

[KLL*24] KARRAS J., LI Y., LIU N., ZHU L., YOO I., LUGMAYR A., LEE C., KEMELMACHER-SHLIZERMAN I.: Fashion-vdm: Video diffusion model for virtual try-on. In *Proceedings of ACM SIGGRAPH Asia 2024* (December 2024).

[KMPA21] KREBS F., MEIXNER A., PATZER I., ASFOUR T.: The kit bimanual manipulation dataset. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)* (2021).

[KNF23] KURIBAYASHI M., NAKAI K., FUNABIKI N.: Image-based virtual try-on system with clothing-size adjustment, 2023.

[KPBD19] KOLOTOUROS N., PAVLAKOS G., BLACK M. J., DANIILIDIS K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV* (2019).

[KPD19] KOLOTOUROS N., PAVLAKOS G., DANIILIDIS K.: Convolutional mesh regression for single-image human shape reconstruction. In *CVPR* (2019).

[KSH23] KOROSTELEVA M., SORKINE-HORNUNG O.: GarmentCode: Programming parametric sewing patterns. *ACM Transaction on Graphics 42*, 6 (2023). SIGGRAPH ASIA 2023 issue.

[KTZ*24] KONG W., TIAN Q., ZHANG Z., MIN R., DAI Z., ZHOU J., XIONG J., LI X., WU B., ZHANG J., ET AL.: Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).

[Kv05] KAVAN L., ŽÁRA J.: Spherical blend skinning: a real-time deformation of articulated models. In *Proceedings of the 2005 Symposium on Interactive 3D Graphics and Games* (New York, NY, USA, 2005), I3D '05, Association for Computing Machinery, p. 9–16.

[KZB*24] KABADAYI B., ZIELONKA W., BHATNAGAR B. L., PONS-MOLL G., THIES J.: Gan-avatar: Controllable personalized gan-based human head avatar. In *International Conference on 3D Vision (3DV)* (March 2024).

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (2017), 194:1–194:17.

[LC87] LORENSEN W. E., CLINE H. E.: Marching cubes: A high resolution 3D surface construction algorithm. *COMPUTER GRAPHICS 21*, 4 (1987), 163–169.

[LCF00] LEWIS J., CORDNER M., FONG N.: Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation. vol. 6, pp. 165–172.

[LCT18a] LAEHNER Z., CREMERS D., TUNG T.: Deepwrinkles: Accurate and realistic clothing modeling. In *European Conference on Computer Vision (ECCV)* (Sept. 2018).

[LCT18b] LÄHNER Z., CREMERS D., TUNG T.: Deepwrinkles: Accurate and realistic clothing modeling. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV* (Berlin, Heidelberg, 2018), Springer-Verlag, p. 698–715.

[LFB17] LEROY V., FRANCO J., BOYER E.: Multi-view dynamic shape refinement using local temporal integration. In *IEEE International Conference on Computer Vision, ICCV* (Venice, Italy, Oct. 2017), pp. 3113–3122.

[LFC*25] LI B., FENG H., CAI Z., BLACK M. J., XIU Y.: Etch: Generalizing body fitting to clothed humans via equivariant tightness. *arXiv preprint arXiv:2503.10624* (2025).

[LGF23] LI R., GUILLARD B., FUA P.: ISP: Multi-Layered Garment Draping with Implicit Sewing Patterns. In *Advances in Neural Information Processing Systems* (2023).

[LGHT24] LIAO Z., GOLYANIK V., HABERMANN M., THEOBALT C.: Vinecs: Video-based neural character skinning. In *Computer Vision and Pattern Recognition (CVPR)* (2024).

[LH21a] LUO S., HU W.: Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021).

[LH21b] LUO S., HU W.: Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021).

[LIPM19a] LAZOVA V., INSAFUTDINOV E., PONS-MOLL G.: 360-degree textures of people in clothing from a single image. In *International Conference on 3D Vision (3DV)* (Sept. 2019).

[LIPM19b] LAZOVA V., INSAFUTDINOV E., PONS-MOLL G.: 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)* (2019), IEEE, pp. 643–653.

[LKL23] LEE D., KANG H., LEE I.-K.: Clothcombo: Modeling inter-cloth interaction for draping multi-layered clothes. *ACM Trans. Graph. 42*, 6 (Dec. 2023).

[LLS*24] LIN S., LI Z., SU Z., ZHENG Z., ZHANG H., LIU Y.: Layga: Layered gaussian avatars for animatable clothing transfer, 2024.

[LLW25] LI J., LIU K. C., WU J.: Mvlift: Towards lifting multi-view images to 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025).

[LMB*14a] LIN T.-Y., MAIRE M., BELONGIE S., BOURDEV L., GIRSHICK R., HAYS J., PERONA P., RAMANAN D., ZITNICK C. L., DOLLÁR P.: Microsoft COCO: Common Objects in Context, 2014.

[LMB14b] LOPER M. M., MAHMOOD N., BLACK M. J.: MoSh: Motion and shape capture from sparse markers. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* (2014).

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia) 34*, 6 (2015), 248:1–248:16.

[LOZ*24] LUO H., OUYANG M., ZHAO Z., JIANG S., ZHANG L., ZHANG Q., YANG W., XU L., YU J.: Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483* (2024).

[LSZ*24] LI X., SUN Q. C., ZHANG P., YE F., LIAO Z., FENG W., ZHAO S., HE Q.: Anydressing: Customizable multi-garment virtual dressing via latent diffusion models. *arXiv preprint arXiv:2412.04146* (2024).

[LTLH19] LIU Z., TANG H., LIN Y., HAN S.: *Point-voxel CNN for efficient 3D deep learning*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[LWP*23] LEI J., WANG Y., PAVLAKOS G., LIU L., DANIILIDIS K.: Gart: Gaussian articulated template models, 2023.

[LYW25] LI X., YAO Q., WANG Y.: Garmentdiffusion: 3d garment sewing pattern generation with multimodal diffusion transformers, 2025.

[LYXC24] LI M., YAO S., XIE Z., CHEN K.: Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024.

[LZL*23] LIN J., ZENG A., LU S., CAI Y., ZHANG R., WANG H., ZHANG L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems* (2023).

[LZT*19] LIU L., ZHENG Y., TANG D., YUAN Y., FAN C., ZHOU K.: Neuroskinning: Automatic skin binding for production characters with deep graph networks. *ACM Trans. Graph. 38*, 4 (July 2019).

[LZWL24] LI Z., ZHENG Z., WANG L., LIU Y.: Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[MBT*12] MIGUEL E., BRADLEY D., THOMASZEWSKI B., BICKEL B., MATUSIK W., OTADUY M. A., MARSCHNER S.: Data-driven estimation of cloth simulation models. *Comput. Graph. Forum 31*, 2 (2012), 519–528.

[MC10] MÜLLER M., CHENTANEZ N.: Wrinkle Meshes. In *Eurographics/ ACM SIGGRAPH Symposium on Computer Animation* (2010), Popovic M., Otaduy M., (Eds.), The Eurographics Association.

[MCC11] MCLAUGHLIN T., CUTLER L., COLEMAN D.: Character rigging, deformations, and simulations in film and game production.

[MCPM24] MARIN R., CORONA E., PONS-MOLL G.: Nicp: Neural icp for 3d human registration at scale. In *European Conference on Computer Vision* (Oct. 2024).

[MGT*19a] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of motion capture as surface shapes. In *ICCV* (2019).

[MGT*19b] MAHMOOD N., GHORBANI N., TROJE N. F., PONS-MOLL G., BLACK M. J.: AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision* (Oct. 2019), pp. 5442–5451.

[MMC22] MUSONI P., MELZI S., CASTELLANI U.: GIM3D: A 3d dataset for garment segmentation. *placeholder* (2022).

[MMC23] MUSONI P., MELZI S., CASTELLANI U.: Gim3d plus: A labeled 3d dataset to design data-driven solutions for dressed humans. *Graphical Models 129* (2023), 101187.

[MNTT12] MATSUYAMA T., NOBUHARA S., TAKAI T., TUNG T.: *3D Video and Its Applications*. Springer, 2012.

[MON*19a] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of CVPR* (2019).

[MON*19b] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy networks: Learning 3D reconstruction in function space. In *CVPR* (2019).

[MSB*22] MIHAJLOVIC M., SAITO S., BANSAL A., ZOLLHOEFER M., TANG S.: COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2022).

[MSJG16] MAKHZANI A., SHLENS J., JAITLY N., GOODFELLOW I.: Adversarial autoencoders. In *ICLR* (2016).

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV* (2020).

[MYBT22] MA Q., YANG J., BLACK M. J., TANG S.: Neural point-based shape modeling of humans in challenging clothing. In *International Conference on 3D Vision (3DV)* (Sept. 2022), pp. 679–689.

[MYCL24] MEN Y., YAO Y., CUI M., LIEFENG B.: Mimo: Controllable character video synthesis with spatial decomposed modeling.

[MYP*24] MÜLLER L., YE V., PAVLAKOS G., BLACK M. J., KANAZAWA A.: Generative proxemics: A prior for 3d social interaction from images. In *CVPR* (2024).

[MYR*20] MA Q., YANG J., RANJAN A., PUJADES S., PONS-MOLL G., TANG S., BLACK M. J.: Learning to Dress 3D People in Generative Clothing. In *Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[MYTB21] MA Q., YANG J., TANG S., BLACK M. J.: The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2021), pp. 10974–10984.

[MZBT21] MIHAJLOVIC M., ZHANG Y., BLACK M. J., TANG S.: LEAP: Learning articulated occupancy of people. In *CVPR* (2021).

[NAK*25] NAKAYAMA K., ACKERMANN J., KESDOGAN T. L., ZHENG Y., KOROSTELEVA M., SORKINE-HORNUNG O., GUIBAS L., YANG G., WETZSTEIN G.: Aipparel: A large multimodal generative model for digital garments. *Computer Vision and Pattern Recognition (CVPR)* (2025).

[NdSL*22] NOVELLO T., DA SILVA V., LOPES H., SCHARDONG G., SCHIRMER L., VELHO L.: Neural implicit surfaces in higher dimension, 2022.

[OBB20] OSMAN A. A. A., BOLKART T., BLACK M. J.: STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)* (2020), pp. 598–613.

[OLPM*18] OMRAN M., LASSNER C., PONS-MOLL G., GEHLER P., SCHIELE B.: Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV* (2018).

[OSBH00] ORMONEIT D., SIDENBLADH H., BLACK M., HASTIE T.: Learning and tracking cyclic human motion. *Advances in Neural Information Processing Systems 13* (2000).

[PBV21] PETROVICH M., BLACK M. J., VAROL G.: Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV* (2021).

[PCG*19] PAVLAKOS G., CHOUTAS V., GHORBANI N., BOLKART T., OSMAN A. A. A., TZIONAS D., BLACK M. J.: Expressive body capture: 3D hands, face, and body from a single image. In *CVPR* (2019).

[PCPMMN20] PUMAROLA A., CORONA E., PONS-MOLL G., MORENO-NOGUER F.: D-NeRF: Neural radiance fields for dynamic scenes. In *CVPR* (2020).

[PDW*21] PENG S., DONG J., WANG Q., ZHANG S., SHUAI Q., ZHOU X., BAO H.: Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV* (2021).

[PFS*19a] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 165–174.

[PFS*19b] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR* (2019).

[PLC*24] PANG H. E., LIU S., CAI Z., YANG L., ZHANG T., LIU Z.: Disco4d: Disentangled 4d human generation and animation from a single image, 2024.

[PLPM20] PATEL C., LIAO Z., PONS-MOLL G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020), IEEE.

[PMPHB17a] PONS-MOLL G., PUJADES S., HU S., BLACK M.: Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH) 36*, 4 (2017). Two first authors contributed equally.

[PMPHB17b] PONS-MOLL G., PUJADES S., HU S., BLACK M.: ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics 36*, 4 (2017).

[PMR*23] PROKUDIN S., MA Q., RAAFAT M., VALENTIN J., TANG S.: Dynamic point fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2023), pp. 7964–7976.

[PMRMB15] PONS-MOLL G., ROMERO J., MAHMOOD N., BLACK M. J.: Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics 34* (2015).

[PSB*20] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Deformable neural radiance fields. *arXiv preprint arXiv:2011.12948* (2020).

[PSC*19] PUMAROLA A., SANCHEZ J., CHOI G., SANFELIU A., MORENO-NOGUER F.: 3DPeople: Modeling the Geometry of Dressed Humans. In *International Conference in Computer Vision (ICCV)* (2019).

[PZK*24] PANG H., ZHU H., KORTYLEWSKI A., THEOBALT C., HABERMANN M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024), pp. 1165–1175.

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. *Proceedings of CVPR* (2017).

[QWM*24] QIAN Z., WANG S., MIHAJLOVIC M., GEIGER A., TANG S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting.

[QXL*22] QIAN S., XU J., LIU Z., MA L., GAO S.: Unif: United neural implicit functions for clothed human reconstruction and animation. In *European Conference on Computer Vision* (2022), Springer, pp. 121–137.

[QYSG17a] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems (NeurIPS)* (2017).

[QYSG17b] QI C. R., YI L., SU H., GUIBAS L. J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2017), NIPS'17, Curran Associates Inc., p. 5105–5114.

[Ras25] RASPBERRY AI: Generative ai for fashion creatives, 2025. Accessed: 2025-06-10. URL: https://www.raspberry.ai/.

[RASS24] RIVERO A., ATHAR S., SHU Z., SAMARAS D.: Rig3dgs: Creating controllable portraits from casual monocular videos, 2024.

[RBD*02] ROBINETTE K., BLACKWELL S., DAANEN H., BOEHMER M., S F., BRILL T., HOEFERLIN D., , D. B.: Civilian american and european surface anthropometry resource (caesar) final report, 2002.

[RBH*21] REMPE D., BIRDAL T., HERTZMANN A., YANG J., SRIDHAR S., GUIBAS L. J.: HuMoR: 3D human motion model for robust pose estimation. In *ICCV* (2021).

[RBL*21] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models, 2021.

[ren] 3D People from Renderpeople. URL: https://renderpeople.com/3d-people/.

[reta] Ai is slashing fashion ecommerce. URL: https://mirrorsize.com/blogs/reduce-e-commerce-returns-in-fashion-industry-with-ai.

[retb] The high cost of retail returns. URL: https://www.thebalancesmb.com/the-high-cost-of-retail-returns-2890350.

[RGRF22] REN L., GUILLARD B., REMELLI E., FUA P.: DIG: Draping Implicit Garment over the Human Body. In *Asian Conference on Computer Vision* (2022).

[RGW*25] RONG B., GRIGOREV A., WANG W., BLACK M. J., THOMASZEWSKI B., TSALICOGLOU C., HILLIGES O.: Gaussian Garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. In *International Conference on 3D Vision 2025* (2025).

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: Grabcut: Interactive foreground extraction using iterated graph cuts. vol. 23.

[RKP*07] ROSENHAHN B., KERSTING U., POWELL K., KLETTE R., KLETTE G., SEIDEL H.-P.: A system for articulated tracking incorporating a clothing model, 2007.

[RLGCC23] ROSS B. L., LOAIZA-GANEM G., CATERINI A. L., CRESSWELL J. C.: Neural implicit manifold learning for topology-aware density estimation. *Transactions on Machine Learning Research* (2023). Expert Certification.

[RPC*10] ROHMER D., POPA T., CANI M.-P., HAHMANN S., ALLA S.: Animation Wrinkling: Augmenting Coarse Cloth Simulations with Realistic-Looking Wrinkles. *ACM Transactions on Graphics 29*, 5 (Dec. 2010), 0. Proceedings of ACM SIGGRAPH ASIA.

[RRN*20] RAVI N., REIZENSTEIN J., NOVOTNY D., GORDON T., LO W.-Y., JOHNSON J., GKIOXARI G.: Accelerating 3D deep learning with PyTorch3D. *arXiv:2007.08501* (2020).

[RTB17] ROMERO J., TZIONAS D., BLACK M. J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 36*, 6 (Nov. 2017).

[Run24] RUNWAY: 48 hours to generate a film. https://runwayml.com/customers/48-hours-to-generate-a-film, 2024. Accessed: 2025-06-10. URL: https://runwayml.com/customers/48-hours-to-generate-a-film.

[SBB10] SIGAL L., BALAN A., BLACK M. J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision 87*, 1 (Mar. 2010), 4–27.

[SBF00] SIDENBLADH H., BLACK M. J., , FLEET D.: Stochastic tracking of 3D human figures using 2D image motion. In *ECCV* (2000).

[SBIK16] SARAFIANOS N., BOTEANU B., IONESCU B., KAKADIARIS I. A.: 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding 152* (2016), 1–20.

[SCD*25] SUN M., CHEN J., DONG J., CHEN Y., JIANG X., MAO S., JIANG P., WANG J., DAI B., HUANG R.: Drive: Diffusion-based rigging empowers generation of versatile and expressive characters.

[SDK*25] SUZUKI K., DU B., KRISHNAN G., CHEN K., LI R. B., NGUYEN T.: Open-vocabulary semantic part segmentation of 3d human, March 2025.

[SDL*25] SETH S., DABRAL R., LUVIZON D., HABERMANN M., YANG M.-H., THEOBALT C., KORTYLEWSKI A.: Pocoloco: A point cloud diffusion model of human shape in loose clothing. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2025).

[SG18] STUTZ D., GEIGER A.: Learning 3d shape completion under weak supervision. *CoRR abs/1805.07290* (2018).

[SGdA*10] STOLL C., GALL J., DE AGUIAR E., THRUN S., THEOBALT C.: Video-based reconstruction of animatable human characters. In *ACM SIGGRAPH Asia* (2010).

[SGK*23] SHEN K., GUO C., KAUFMANN M., ZARATE J., VALENTIN J., SONG J., HILLIGES O.: X-avatar: Expressive human avatars. *Computer Vision and Pattern Recognition (CVPR)* (2023).

[SGY*21] SHEN T., GAO J., YIN K., LIU M.-Y., FIDLER S.: Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

[SH07] STARCK J., HILTON A.: Surface capture for performance-based animation. *IEEE Computer Graphics and Applications 27*, 3 (2007), 21–31.

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 2304–2314.

[SLD23] SHAO Y., LOY C. C., DAI B.: Towards multi-layered 3D garments animation. In *ICCV* (2023).

[SLNW16] SHAHROUDY A., LIU J., NG T.-T., WANG G.: NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR* (2016).

[SMAB24] SVITOV D., MORERIO P., AGAPITO L., BUE A. D.: Haha: Highly articulated gaussian human avatars with textured mesh prior, 2024.

[SMB*20] SITZMANN V., MARTEL J. N., BERGMAN A. W., LINDELL D. B., WETZSTEIN G.: Implicit neural representations with periodic activation functions. In *NeurIPS* (2020).

[SMR*24] SRIVASTAVA A., MANU P., RAJ A., JAMPANI V., SHARMA A.: Wordrobe: Text-guided generation of textured 3d garments, 2024.

[SNTH03] SHAO W., NG-THOW-HING V.: A general joint component framework for realistic articulation in human characters. In *Proceedings of the 2003 symposium on Interactive 3D graphics* (2003), pp. 11–18.

[SOC19] SANTESTEBAN I., OTADUY M. A., CASAS D.: Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)* (2019).

[SOC22] SANTESTEBAN I., OTADUY M. A., CASAS D.: SNUG: Self-Supervised Neural Dynamic Garments. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[SOTC22] SANTESTEBAN I., OTADUY M. A., THUEREY N., CASAS D.: ULNeF: Untangled layered neural fields for mix-and-match virtual try-on. In *Advances in Neural Information Processing Systems, (NeurIPS)* (2022).

[SP91] SCLAROFF S., PENTLAND A.: Generalized implicit functions for computer graphics. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1991), SIGGRAPH '91, Association for Computing Machinery, p. 247–250.

[SPCM97] SCHEEPERS F., PARENT R. E., CARLSON W. E., MAY S. F.: Anatomy-based modeling of the human musculature. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques* (USA, 1997), SIGGRAPH '97, ACM Press/Addison-Wesley Publishing Co., p. 163–172.

[SPK19] SHU D. W., PARK S. W., KWON J.: 3d point cloud generative adversarial network based on tree structured graph convolutions. *CoRR abs/1905.06292* (2019).

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (June 2020).

[ST03] SEO H., THALMANN N.: An automatic modeling of human bodies from sizing parameters. pp. 19–26.

[Stu18] STUYCK T.: *Cloth Simulation for Computer Graphics*. Synthesis Lectures on Visual Computing. Morgan & Claypool Publishers, 2018.

[SWL*24] SHAO Z., WANG Z., LI Z., WANG D., LIN X., ZHANG Y., FAN M., WANG Z.: SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[SYMB21] SAITO S., YANG J., MA Q., BLACK M. J.: SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (June 2021).

[SYWL23] SU Z., YU T., WANG Y., LIU Y.: Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 2 (2023), 1581–1593.

[SZZ*22] SHAO R., ZHENG Z., ZHANG H., SUN J., LIU Y.: Diffustereo: High quality human reconstruction via diffusion-based stereo using sparse cameras. In *ECCV* (2022).

[TAL*22] TIWARI G., ANTIC D., LENSSEN J. E., SARAFIANOS N., TUNG T., PONS-MOLL G.: Pose-ndf: Modeling human pose manifolds with neural distance fields. In *European Conference on Computer Vision (ECCV)* (Oct. 2022), Springer.

[TBTPM20] TIWARI G., BHATNAGAR B. L., TUNG T., PONS-MOLL G.: SIZER: A dataset and model for parsing 3D clothing and learning size sensitive 3D clothing. In *ECCV* (2020).

[TCC*24] TANG J., CHEN Z., CHEN X., WANG T., ZENG G., LIU Z.: LGM: large multi-view gaussian model for high-resolution 3d content creation. *CoRR abs/2402.05054* (2024).

[TDK*24] TANG J., DAVOLI D., KIRSCHSTEIN T., SCHONEVELD L., , NIESSNER M.: Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion.

[Tiw20] TIWARI G.: Learning size sensitive cloth model. Computer Vision and Machine Learning, MPI for Informatics, Max Planck Society.

[TNM08] TUNG T., NOBUHARA S., MATSUYAMA T.: Simultaneous super-resolution and 3d video using graph-cuts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2008).

[TNM09] TUNG T., NOBUHARA S., MATSUYAMA T.: Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE 12th International Conference on Computer Vision, ICCV* (Kyoto, Japan, Sept. 2009), pp. 1709–1716.

[TQD*19] THOMAS H., QI C. R., DESCHAUD J.-E., MARCOTEGUI B., GOULETTE F., GUIBAS L. J.: Kpconv: Flexible and deformable convolution for point clouds. *Proceedings of the IEEE International Conference on Computer Vision* (2019).

[tre] Treedy's scanner. URL: https://www.treedys.com.

[TRG*23] TEVET G., RAAB S., GORDON B., SHAFIR Y., COHEN-OR D., BERMANO A. H.: Human motion diffusion model. In *The Eleventh International Conference on Learning Representations* (2023).

[Tru25] TRUE FIT CORPORATION: True fit – personalization platform for apparel and footwear, 2025. Accessed: 2025-05-06. URL: https://www.truefit.com/.

[TSTPM21] TIWARI G., SARAFIANOS N., TUNG T., PONS-MOLL G.: Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *ICCV* (2021).

[TW17] TOMCZAK J. M., WELLING M.: VAE with a VampPrior. *arXiv* (2017).

[twi] Twindom 3D Scans. URL: https://web.twindom.com/.

[TXL*25] TIWARI G., XUE Y., LENSSEN J. E., TUNG T., PONS-MOLL G.: Tailordiff: A unified and expressive generative model for 3d clothing. Manuscript in preparation, 2025.

[TZZ*19] TAO Y., ZHENG Z., ZHONG Y., ZHAO J., QUIONHAI D., PONS-MOLL G., LIU Y.: Simulcap : Single-view human performance capture with cloth simulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019).

[UFF06] URTASUN R., FLEET D., FUA P.: 3D people tracking with Gaussian process dynamical models. In *CVPR* (2006).

[VCMT95] VOLINO P., COURCHESNE M., MAGNENAT THALMANN N.: Versatile and efficient techniques for simulating cloth and other deformable objects. In *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 1995), SIGGRAPH '95, Association for Computing Machinery, p. 137–144.

[VKK21] VAHDAT A., KREIS K., KAUTZ J.: Score-based generative modeling in latent space. In *Neural Information Processing Systems (NeurIPS)* (2021).

[vMHB*18a] VON MARCARD T., HENSCHEL R., BLACK M., ROSENHAHN B., PONS-MOLL G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In *ECCV* (2018).

[vMHB*18b] VON MARCARD T., HENSCHEL R., BLACK M. J., ROSENHAHN B., PONS-MOLL G.: Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV* (2018).

[VSGC20] VIDAURRE R., SANTESTEBAN I., GARCES E., CASAS D.: Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)* (2020).

[VSJ22] VICINI D., SPEIERER S., JAKOB W.: Differentiable signed distance function rendering. *Transactions on Graphics (Proceedings of SIGGRAPH) 41*, 4 (July 2022), 125:1–125:18. doi:10.1145/3528223.3530139.

[VT94] VOLINO P., THALMANN N. M.: Efficient self-collision detection on smoothly discretized surface animations using geometrical shape regularity. *Computer Graphics Forum 13*, 3 (1994), 155–166.

[WAW*25] WANG A., AI B., WEN B., MAO C., XIE C.-W., CHEN D., YU F., ZHAO H., YANG J., ZENG J., WANG J., ZHANG J., ZHOU J., WANG J., CHEN J., ZHU K., ZHAO K., YAN K., HUANG L., FENG M., ZHANG N., LI P., WU P., CHU R., FENG R., ZHANG S., SUN S., FANG T., WANG T., GUI T., WENG T., SHEN T., LIN W., WANG W., WANG W., ZHOU W., WANG W., SHEN W., YU W., SHI X., HUANG X., XU X., KOU Y., LV Y., LI Y., LIU Y., WANG Y., ZHANG Y., HUANG Y., LI Y., WU Y., LIU Y., PAN Y., ZHENG Y., HONG Y., SHI Y., FENG Y., JIANG Z., HAN Z., WU Z.-F., LIU Z.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025).

[WCF07] WHITE R., CRANE K., FORSYTH D. A.: Capturing and animating occluded cloth. *ACM Trans. Graph. 26*, 3 (2007), 34.

[WCPM18] WANG T. Y., CEYLAN D., POPOVIC J., MITRA N. J.: Learning a shared shape space for multimodal garment design. *ACM Trans. Graph. 37*, 6 (2018), 1:1–1:14.

[WCS*22] WENG C., CURLESS B., SRINIVASAN P. P., BARRON J. T., KEMELMACHER-SHLIZERMAN I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, jun 18-24, 2022* (2022), IEEE, pp. 16189–16199.

[WHG*24] WANG W., HO H.-I., GUO C., RONG B., GRIGOREV A., SONG J., ZARATE J. J., HILLIGES O.: 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[WHH*23] WANG Y., HAN Q., HABERMANN M., DANIILIDIS K., THEOBALT C., LIU L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).

[WHRO10] WANG H., HECHT F., RAMAMOORTHI R., O'BRIEN J. F.: Example-based wrinkle synthesis for clothing animation. *ACM Transactions on Graphics (Proceedings of SIGGRAPH) 29*, 4 (July 2010), 107:1–8.

[WLD*23] WANG J., LIU Y., DOU Z., YU Z., LIANG Y., LI X., WANG W., XIE R., SONG L.: Disentangled clothed avatar generation from text descriptions, 2023.

[WNEH22] WIERSMA R., NASIKUN A., EISEMANN E., HILDEBRANDT K.: Deltaconv: Anisotropic operators for geometric deep learning on point clouds. *Transactions on Graphics 41*, 4 (July 2022).

[WRO11] WANG H., RAMAMOORTHI R., O'BRIEN J. F.: Data-driven elastic models for cloth: Modeling and measurement. *ACM Transactions on Graphics (Proceedings of SIGGRAPH) 30*, 4 (July 2011), 71:1–11.

[WSFM] WANG Y., SHAO T., FU K., MITRA N.: Learning an intrinsic garment space for interactive authoring of garment animation. *ACM Trans. Graph. 38*, 6.

[WSGT22] WANG S., SCHWARZ K., GEIGER A., TANG S.: Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision* (2022).

[WSL*19] WANG Y., SUN Y., LIU Z., SARMA S. E., BRONSTEIN M. M., SOLOMON J. M.: Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* (2019).

[WYHN18] WANG W., YU R., HUANG Q., NEUMANN U.: Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *CVPR* (2018).

[XBLPM24] XIE X., BHATNAGAR B. L., LENSSEN J. E., PONS-MOLL G.: Template free reconstruction of human-object interaction with procedural interaction generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2024).

[XBPM22] XIE X., BHATNAGAR B. L., PONS-MOLL G.: Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)* (Oct. 2022), Springer.

[XJS19] XIANG D., JOO H., SHEIKH Y.: Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 10965–10974.

[XLL*23] XIONG Z., LI C., LIU K., LIAO H., HU J., ZHU J., NING S., QIU L., WANG C., WANG S., CUI S., HAN X.: Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. *CoRR abs/2312.02963* (2023).

[XLPM25] XIE X., LENSSEN J. E., PONS-MOLL G.: Intertrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)* (March 2025).

[XTS*22] XIE Y., TAKIKAWA T., SAITO S., LITANY O., YAN S., KHAN N., TOMBARI F., TOMPKIN J., SITZMANN V., SRIDHAR S.: Neural fields in visual computing and beyond. *Computer Graphics Forum* (2022).

[XXMPM24] XUE Y., XIE X., MARIN R., PONS-MOLL G.: Human 3diffusion: Realistic avatar creation via explicit 3d consistent diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)* (Dec. 2024).

[XYC*23] XIU Y., YANG J., CAO X., TZIONAS D., BLACK M. J.: ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2023).

[XYL*24] XIU Y., YE Y., LIU Z., TZIONAS D., BLACK M. J.: Puzzleavatar: Assembling 3d avatars from personal albums. *ACM Transactions on Graphics (TOG)* (2024).

[XYTB22] XIU Y., YANG J., TZIONAS D., BLACK M. J.: ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), pp. 13296–13306.

[XZK*20] XU Z., ZHOU Y., KALOGERAKIS E., LANDRETH C., SINGH K.: Rignet: Neural rigging for articulated characters. *ACM Trans. on Graphics 39* (2020).

[XZT19] XU Y., ZHU S.-C., TUNG T.: Denserac: Joint 3D pose and shape estimation by dense render and compare. In *International Conference on Computer Vision* (2019).

[Yam12] YAMAGUCHI K.: Parsing clothing in fashion photographs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (USA, 2012), CVPR '12, IEEE Computer Society, p. 3570–3577.

[YFHWW18] YANG J., FRANCO J.-S., HÉTROY-WHEELER F., WUHRER S.: Analyzing clothing layer deformation statistics of 3D human motions. In *Computer Vision – ECCV 2018* (Cham, 2018), Ferrari V., Hebert M., Sminchisescu C., Weiss Y., (Eds.), Springer International Publishing, pp. 245–261.

[YHH*19] YANG G., HUANG X., HAO Z., LIU M.-Y., BELONGIE S., HARIHARAN B.: Pointflow: 3d point cloud generation with continuous normalizing flows. *arXiv* (2019).

[YKB13] YAMAGUCHI K., KIAPOUR M. H., BERG T. L.: Paper doll parsing: Retrieving similar styles to parse clothing items. In *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013* (2013), IEEE Computer Society, pp. 3519–3526.

[YL14] YANG W., LUO P.AND LIN L.: Clothing co-parsing by joint image segmentation and labeling.

[YNU24] YAMASHITA Y., NAKATANI C., UKITA N.: Size-variable virtual try-on with physical clothes size, 2024.

[YTB*21] YENAMANDRA T., TEWARI A., BERNARD F., SEIDEL H., ELGHARIB M., CREMERS D., THEOBALT C.: i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021).

[YWM*21] YANG Z., WANG S., MANIVASAGAM S., HUANG Z., MA W.-C., YAN X., YUMER E., URTASUN R.: S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR* (2021).

[YZG*18] YU T., ZHENG Z., GUO K., ZHAO J., DAI Q., LI H., PONS-MOLL G., LIU Y.: Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)* (June 2018), IEEE.

[YZG*21] YU T., ZHENG Z., GUO K., LIU P., DAI Q., LIU Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)* (June 2021).

[ZBLPM22] ZHOU K., BHATNAGAR B. L., LENSSEN J. E., PONS-MOLL G.: Toch: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)* (Oct. 2022), Springer.

[ZBS*25] ZIELONKA W., BAGAUTDINOV T., SAITO S., ZOLLHÖFER M., THIES J., ROMERO J.: Drivable 3d gaussian avatars. In *International Conference on 3D Vision (3DV)* (Mar. 2025).

[ZCJ*20] ZHU H., CAO Y., JIN H., CHEN W., DU D., WANG Z., CUI S., HAN X.: Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images, 2020.

[ZDW21] ZHOU L., DU Y., WU J.: 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2021), pp. 5826–5835.

[ZGV*25] ZHOU J. J., GAO H., VOLETI V., VASISHTA A., YAO C.-H., BOSS M., TORR P., RUPPRECHT C., JAMPANI V.: Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint* (2025).

[ZHN*20] ZHANG Y., HASSAN M., NEUMANN H., BLACK M. J., TANG S.: Generating 3D people in scenes without people. In *CVPR* (2020).

[ZHW23] ZOU X., HAN X., WONG W.: Cloth4d: A dataset for clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12847–12857.

[ZHY*22] ZHENG Z., HUANG H., YU T., ZHANG H., GUO Y., LIU Y.: Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022).

[ZJJ*20] ZHAO H., JIANG L., JIA J., TORR P., KOLTUN V.: Point transformer, 2020.

[ZLL*24] ZHENG X., LIAO L., LI X., JIAO J., WANG R., GAO F., WANG S., WANG R.: Pku-dymvhumans: A multi-view video benchmark for high-fidelity dynamic human modeling. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2024).

[ZLZ*25] ZHANG Y., LIN J., ZENG A., WU G., LU S., FU Y., CAI Y., ZHANG R., WANG H., ZHANG L.: Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. *arXiv preprint arXiv:2501.05098* (2025).

[ZMGL21] ZAKHARKIN I., MAZUR K., GRIGOREV A., LEMPITSKY V.: Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 14718–14727.

[ZPBP17] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017* (2017), IEEE Computer Society, pp. 5484–5493.

[ZPBPM17a] ZHANG C., PUJADES S., BLACK M., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE CVPR* (2017).

[ZPBPM17b] ZHANG C., PUJADES S., BLACK M. J., PONS-MOLL G.: Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017).

[ZPG*25] ZHANG J.-P., PU C.-F., GUO M.-H., CAO Y.-P., HU S.-M.: One model to rig them all: Diverse skeleton rigging with unirig.

[ZSZ*21] ZHENG Y., SHAO R., ZHANG Y., YU T., ZHENG Z., DAI Q., LIU Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *IEEE Conference on Computer Vision (ICCV 2021)* (2021).

[ZVW*22] ZENG X., VAHDAT A., WILLIAMS F., GOJCIC Z., LITANY O., FIDLER S., KREIS K.: Lion: Latent point diffusion models for 3d shape generation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

[ZWCM20] ZHANG M., WANG T., CEYLAN D., MITRA N. J.: Deep detail enhancement for any garment. *arXiv preprint arXiv:2008.04367* (2020).

[ZYW*19] ZHENG Z., YU T., WEI Y., DAI Q., LIU Y.: Deephuman: 3d human reconstruction from a single image. In *The IEEE International Conference on Computer Vision (ICCV)* (Oct. 2019).

[ZYZ*22] ZHAO F., YANG W., ZHANG J., LIN P., ZHANG Y., YU J., XU L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, jun 18-24, 2022* (2022), IEEE, pp. 7733–7743.

[ZZB*21] ZHANG S., ZHANG H., BOGO F., POLLEFEYS M., TANG S.: Learning motion priors for 4D human body capture in 3D scenes. In *ICCV* (2021).

[ZZL*23] ZHOU B., ZHOU H., LIANG T., YU Q., ZHAO S., ZENG Y., LV J., LUO S., WANG Q., YU X., CHEN H., LU C., SHAO L.: Clothesnet: An information-rich 3d garment model repository with simulated clothes environment. In *IEEE/CVF International Conference on Computer Vision (ICCV)* (2023).

[ZZM*20] ZHANG S., ZHANG Y., MA Q., BLACK M. J., TANG S.: PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV* (2020).

[ZZQ*20] ZOU S., ZUO X., QIAN Y., WANG S., XU C., GONG M., CHENG L.: 3D human shape reconstruction from a polarization image. In *ECCV* (2020).

[ZZZ*23] ZHENG Z., ZHAO X., ZHANG H., LIU B., LIU Y.: Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG) 42*, 4 (2023).