Towards Disentangled Representation Learning in Practice

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Yash Sharma
aus New York, USA

Tübingen
2024

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

Tag der mündlichen Qualifikation:	16.09.2024
Dekan:	Prof. Dr. Thilo Stehle
Berichterstatter::	Dr. Wieland Brendel
Berichterstatter:	Prof. Dr. Andreas Geiger

Abstract

While the success of deep learning is underpinned by learning representations of data, what information the learned representations extract remains a mystery. In our first contribution (C1), we show that state-of-the-art approaches to self-supervised visual representation learning extract the aspects, or factors of variation (FoVs), of the data that are invariant to data augmentations applied during training, discarding the variant FoVs. In studying augmentations used in practice, we find that while object class is left invariant, position, hue, and rotation information tend to be discarded, which is problematic for tasks outside of object recognition, e.g. object localization. In our second contribution (C2), we show that such approaches can yield *disentangled* representations, where all FoVs are extracted separately in the representation, if all FoVs are variant to the augmentations, an assumption that notably isn't met by augmentations used in practice. In our third contribution (C3), we show evidence that this assumption can be met in natural video, where FoVs undergo transitions that are typically small in magnitude with occasional large jumps, characteristic of a temporally sparse distribution. While challenges remain for real-world disentanglement, our contributions provide guidance to the field in the pursuit of progress in representation learning.

Zusammenfassung

Während der Erfolg von Deep Learning durch das Erlernen von Datenrepräsentationen untermauert wird, bleibt es unklar, welche Informationen die erlernten Repräsentationen etrahieren. In unserem ersten Beitrag (C1) zeigen wir, dass hochmoderne Ansätze zum selbstüberwachten visuellen Repräsentationslernen die Aspekte, oder Variationsfaktoren (FoVs), der Daten extrahieren, die gegenüber der während des Trainings angewendeten Datenaugmentation invariant sind , wobei die varianten FoVs verworfen werden. Bei der Untersuchung von in der Praxis verwendeten Augmentationen stellen wir fest, dass die Objektklasse zwar invariant bleibt, Positions-, Farbton- und Rotationsinformationen jedoch tendenziell verworfen werden, was für Aufgaben außerhalb der Objekterkennung problematisch ist, z. B. Objektlokalisierung. In unserem zweiten Beitrag (C2) zeigen wir, dass solche Ansätze zu disentangled Repräsentationen führen können, bei denen alle FoVs separat in der Darstellung extrahiert werden, wenn alle FoVs variant gegenüber der Augmentation sind, eine Annahme, die insbesondere bei in der Praxis verwendeten Augmentationen nicht erfüllt wird. In unserem dritten Beitrag (C3) zeigen wir Beweise dafür, dass diese Annahme in natürlichen Videos erfüllt werden kann, wo FoVs Übergänge durchlaufen von typischerweise geringer Größenordnung mit gelegentlich großen Sprüngen durchlaufen, was charakteristisch für eine zeitlich dünn besetzte Verteilung ist. Auch wenn Disentanglement in der realen Welt noch vor Herausforderungen steht bieten unsere Beiträge Orientierung für weitere Arbeiten auf der Suche nach Fortschritten beim Lernen von Repräsentationen.

List of Publications

Publications part of thesis (*‡equal contribution)

David A. Klindt*, Lukas Schott*, Yash Sharma*, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge‡, Dylan Paiton‡. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. *International Conference on Learning Representations*, 2021

Roland S Zimmermann*, Yash Sharma*, Steffen Schneider*, Matthias Bethge‡, Wieland Brendel‡. Contrastive Learning Inverts the Data Generating Process. *International Conference on Machine Learning*, 2021

Julius von Kügelgen*, Yash Sharma*, Luigi Gresele*, Wieland Brendel, Bernhard Schölkopf‡, Michel Besserve‡, Francesco Locatello‡. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. *Advances in Neural Information Processing Systems*, 2021

Publications not part of thesis (*‡equal contribution)

Marissa A. Weis, Kashyap Chitta, Yash Sharma, Wieland Brendel, Matthias Bethge, Andreas Geiger, Alexander S. Ecker. Benchmarking Unsupervised Object Representations for Video Sequences. *Journal of Machine Learning Research*, 2021

Nasim Rahaman, Anirudh Goyal, Muhammad Waleed Gondal, Manuel Wuthrich, Stefan Bauer, Yash Sharma, Yoshua Bengio, Bernhard Schölkopf. Spatially Structured Recurrent Modules. *International Conference on Learning Representations*, 2021

Yilun Du, Shuang Li, Yash Sharma, Joshua Tenenbaum, Igor Mordatch. Unsupervised Learning of Compositional Energy Concepts. *Advances in Neural Information Processing Systems*, 2021

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. *Conference on Causal Learning and Reasoning*, 2022

Yash Sharma, Yi Zhu, Chris Russell, Thomas Brox. Pixel-level Correspondence for Self-Supervised Learning from Video. *ICML Workshop on Pre-training*, 2022

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár‡, Wieland Brendel‡. Jacobian-based Causal Discovery with Nonlinear ICA. *Transactions on Machine Learning Research*, 2023

Jack Brady*, Roland S Zimmermann*, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen‡, Wieland Brendel‡. Provably Learning Object-Centric Representations. *International Conference on Machine Learning*, 2023

Laura Fee Nern, Harsh Raj, Maurice Georgi, Yash Sharma. On Transfer of Adversarial Robustness from Pretraining to Downstream Tasks. *Advances in Neural Information Processing Systems*, 2023

Statement of Author Contributions

David A. Klindt*, Lukas Schott*, Yash Sharma*, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge‡, Dylan Paiton‡. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. *International Conference on Learning Representations*, 2021

MB proposed the idea of temporal sparse coding with deep networks repeatedly to the lab; DAK conceived the idea of the model with input from LS and DP; DAK, LS and YS performed the main experiments; LS and YS respectively designed the simplified natural dataset KITTI-Masks and NaturalSprites with input from DAK, DP and WB; DAK, YS and LS analyzed the statistics of the natural datasets; LS performed an in-depth assessment of how latents are encoded in figures 4, 5 and multiple appendix figures with input from DAK, YS and DP; MB structured and supervised the theoretical analysis of the paper; IU, DAK and WB proved theorem 1; DAK derived the objective function for the SlowVAE with input from IU; LS derived the objective function for the SlowFlow with input from DP; LS and YS performed an indepth comparison to PCL in appendix F; DP compared the metrics in appendix C; YS performed the permutation experiments and the transition prior ablation in appendix F as well as the Δt ablation in appendix G; DAK, DP, YS and LS wrote the manuscript with input from WB, IU and MB.

Roland S Zimmermann*, Yash Sharma*, Steffen Schneider*, Matthias Bethge‡, Wieland Brendel‡. Contrastive Learning Inverts the Data Generating Process. *International Conference on Machine Learning*, 2021

The project was initiated by WB. RSZ, SS and WB jointly derived the theory. RSZ and YS implemented and executed the experiments. The 3DIdent dataset was created by RSZ with feedback from SS, YS, WB and MB. RSZ, YS, SS and WB contributed to the final version of the manuscript.

Julius von Kügelgen*, Yash Sharma*, Luigi Gresele*, Wieland Brendel, Bernhard Schölkopf‡, Michel Besserve‡, Francesco Locatello‡. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. *Advances in Neural Information Processing Systems*, 2021

JvK contributed to the conceptualization, theory, design of experiments, analysis of results, and writing. YS suggested prior work for proving theorem 4.4, designed Causal3DIdent with help from JvK, and led all experimental work. LG suggested studying the problem in connection to the multi-view setting, worked on the causal interpretation of the problem and discussed the experiments and theory (particularly in connection with nonlinear ICA theory) with all co-authors. All co-authors contributed to the final product.

Acknowledgements

Matthias Bethge and Wieland Brendel, thank you for the opportunity to spend my PhD in Tübingen. It was an honor to absorb all that I could from excellent scientists regularly producing high-quality research. Andreas Geiger, each and every interaction with you brought a great deal of value, I'm very grateful for the time you've given me. I am highly appreciative of Wieland for giving me the freedom to explore, and thus enabling me to develop as a researcher. Without a doubt, the diversity of experiences I've been afforded has shaped me greatly, and none of this would have been possible without what Tübingen has provided me.

I've had the privilege to collaborate with colleagues across the research landscape. Yilun Du, Michael Chang, Michael Janner, Sébastien Lachapelle, Julius von Kügelgen, Jack Brady, Luigi Gresele, Francesco Locatello I gained so much from our time working together, it was great to learn from each of you. Roland S Zimmermann, Steffen Schneider it was great to both collaborate and witness your research take shape. David A. Klindt, Lukas Schott, Dylan Paiton, SlowVAE was a blast, thanks for setting the tone for my PhD, and for showing me how sweet teamwork can be. Laura Fee Nern, thanks for the opportunity to advise, I'm glad I was able to be a part of your journey. Finally, Preetum Nakkiran, Naomi Saphra, Ekdeep Singh, Vishaal Udandarao, and Ameya Prabhu, meeting you has greatly shaped my post-PhD trajectory, thanks for the meaningful time together.

Lastly, I'd like to thank my family and friends. On my nomadic journey, I can't thank you enough for giving me a base of support. I can only hope to pay it forward.

Contents

Introduction 1

From object recognition (Krizhevsky et al., 2012) to conversational assistance (Ouyang et al., 2022), progress in artificial intelligence (AI) has undoubtedly been driven by deep learning, where data representations are learned to facilitate the extraction of useful information for solving downstream tasks (Bengio et al., 2013). Assuming the availability of labels, supervised learning provides a straightforward way to ensure data representations capture useful information. However, since data labeling can quickly become prohibitively expensive at scale, modern systems have instead turned to *self-supervised learning* (SSL), where approaches are designed for finding useful representations from unlabeled data (Chen et al., 2020; Brown et al., 2020; Balestriero et al., 2023). While the application of SSL has led to technology with significant implications for society (Brundage et al., 2018; Eloundou et al., 2023; EO, 2023), our understanding of what information the learned representations extract remains limited. For instance, in the traditional supervised setting, deep neural networks were understood to solve object recognition by extracting increasingly abstract features from images, i.e. detecting objects as combinations of object parts, where object parts are composed of identified edges (LeCun et al., 2015; Goodfellow et al., 2016). However, although visualization techniques (Zeiler and Fergus, 2014) tend to highlight object parts in features, by experimenting with texture-shape cue conflicts, e.g. cat shape with elephant texture, it was later shown that networks were in fact biased towards texture (Geirhos et al., 2019). Moreover, the reliability of feature visualization (Erhan et al., 2009; Mordvintsev et al., 2015; Olah et al., 2017) has been called into question (Sixt et al., 2020; Borowski et al., 2021; Geirhos et al., 2023). Given this predicament, are there alternative tools worth considering?

In this thesis, we instead take an *identifiability* perspective (Comon, 1994; Hyvärinen and Pajunen, 1999; Hyvärinen et al., 2023), where we study representation learning to understand what gives rise to the relationship between the extracted representations and the information of interest. If we are interested in object recognition, we may want learned representations that not only preserve object class, but where object class can be recovered easily, e.g. via a linear transform (Chen et al., 2020). However, if we are instead interested in autonomous driving (Vlasic and Boudette, 2016), object localization, where an identified object is located as well (Everingham et al., 2015), is also important for an agent (LeCun, 2022) to make reliable decisions in safety-critical scenarios. This begs the question, does finding useful representations require deciding apriori what information is useful, and how it should be represented? Further, can these decisions be enforced in practice?

To address these questions, we present the following contributions. First, in studying state-of-the-art approaches in visual SSL, we find that the information the learned representations represent is entirely determined by the choice of data augmentations applied during training. Second, we find that if we reconsider the choice of data augmentations, the same visual SSL approaches can yield *disentangled* representations, where not only is all information preserved, but information of interest can be recovered by a simple readout of specific representation components. Third, we find that a shift from data augmentations to natural video can yield disentangled representations, suggesting a path to disentanglement in practice.

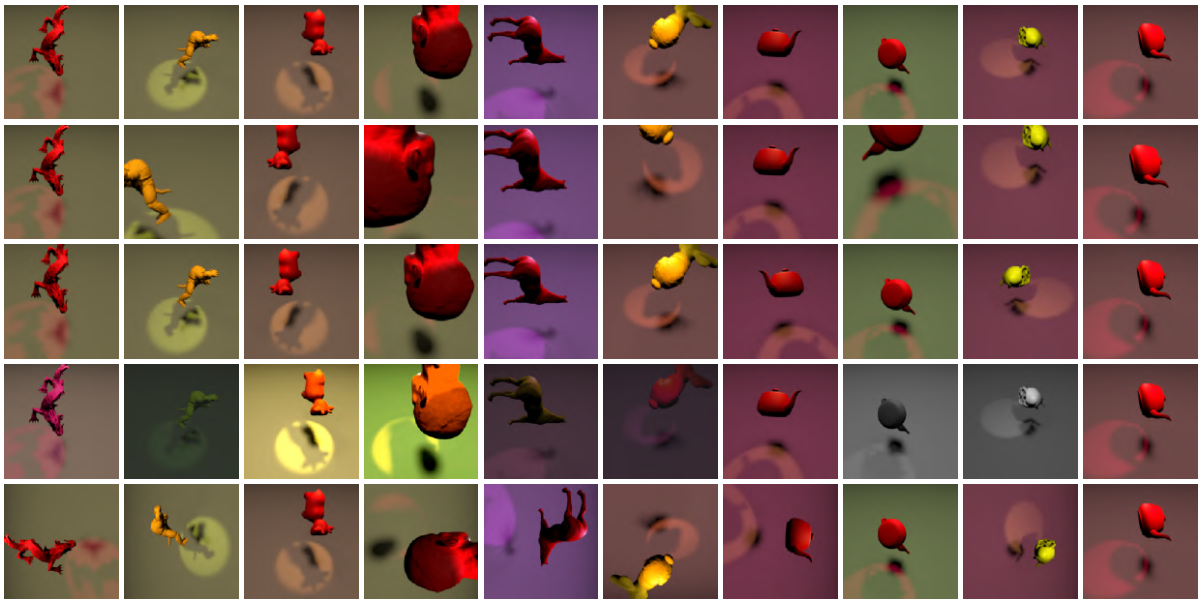


Figure 1.1: Visual overview of the effect of different data augmentations applied to 10 representative samples. Rows correspond to *(top to bottom)*: original images, small random crop (+ random flip), large random crop (+ random flip), color distortion (jitter & drop), and random rotation. Figure from (Von Kügelgen et al., 2021).

1.1 What do representations represent?

For state-of-the-art approaches to SSL, particularly in vision (Oquab et al., 2023), a common practice is to learn representations by maximizing alignment (Wang and Isola, 2020), i.e. similarity, between positive pairs, e.g. data augmentations of the same image, while avoiding collapsed representations through regularization (Chen et al., 2020; Chen and He, 2021; Caron et al., 2021). Contrastive approaches (Hadsell et al., 2006; Dosovitskiy et al., 2014; Oord et al., 2018; Chen et al., 2020) prevent collapse by maximizing dissimilarity between negative pairs, e.g. data augmentations of different images, leading to an instance discrimination objective. Later work found that computing dissimilarity with different images was unnecessary for preventing collapse, as long as augmentations of the same image are processed by asymmetric branches, and gradients for learning are computed using only one of the branches (Grill et al., 2020; Chen and He, 2021; Caron et al., 2021). Finally, non-contrastive approaches can also prevent collapse by maximizing dissimilarity between representation components, thereby reducing redundancy in the representation (Zbontar et al., 2021; Bardes et al., 2022).

In our first contribution (C1) (Von Kügelgen et al., 2021), we show that such approaches learn representations that extract the information invariant to the data augmentations applied during training. Augmentations commonly used in practice are shown in Figure 1.1. While each augmentation can distort some information, there exists information left unchanged across all augmentations, i.e., the object in the original image is depicted in all of the augmented images. Theoretically, we show that representations learned by jointly maximizing alignment and entropy of the representation, thereby avoiding collapse, contains all and only information about *content*, what's preserved across augmentations, while discarding *style*, the rest of the information. Notably, the regularization applied by contrastive approaches has been shown

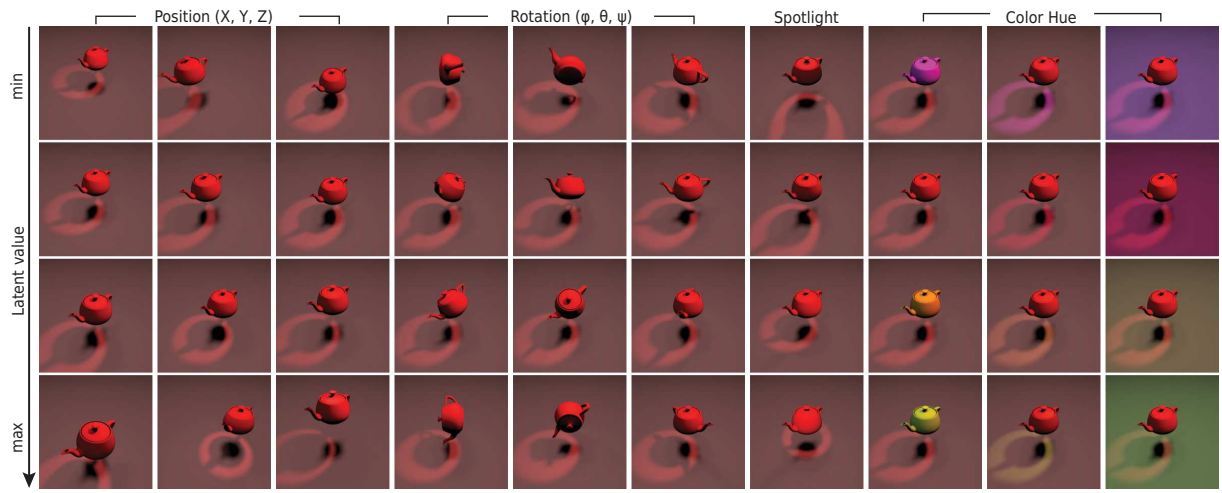


Figure 1.2: Influence of factors of variation (FoVs) on renderings. Each column corresponds to a traversal in one of the 10 latent factors while the other factors are kept fixed. Figure from (Zimmermann et al., 2021) (made by Roland S. Zimmermann).

to maximize entropy (Wang and Isola, 2020), and for non-contrastive approaches, maximizing entropy minimizes redundancy in the representation (Zbontar et al., 2021). Empirically, for contrastive and non-contrastive learning, we find that while we are able to consistently infer object class, the same cannot be said for position, hue, and rotation information.

Given object recognition forms the basis for benchmarking visual SSL, it is unsurprising that state-of-the-art approaches choose data augmentations that will discard irrelevant information for this downstream task. However, as discussed, discarding positional, hue, and rotation information can be problematic when reliable decision-making depends on said variables, e.g. in autonomous driving. If we instead *disentangled* this information, we could maintain state-of-the-art performance on object recognition by isolating object class, while having positional, hue and rotation information separated, such that decision-making can operate on each variable when needed. While recent work has sought to disentangle rather than discard (Eastwood et al., 2023), disentangled representation learning in practice remains an open question.

1.2 When do representations disentangle?

From an identifiability perspective, disentanglement has been studied as a long-standing goal of independent component analysis (ICA) (Comon, 1994). ICA can be viewed as an extension of principal component analysis (PCA), in that it extracts independent as opposed to merely uncorrelated components, enabling identifiability of the latent sources, or the factors of variation (FoVs) (Bengio et al., 2013), underlying the observational data. As can be seen in Figure 1.2, each FoV is information that can be independently varied; while position, rotation, and hue FoVs are depicted for a teapot, varying object identity would result in an object class FoV as well. Disentanglement ensures that each component of the representation isolates different FoVs, e.g. object identity and location are extracted separately.

In our second contribution (C2) (Zimmermann et al., 2021), we show that the aforementioned state-of-the-art visual SSL approaches can in fact achieve disentanglement, if certain statistical assumptions are fulfilled by the data generating process. Empirically, we find that the result

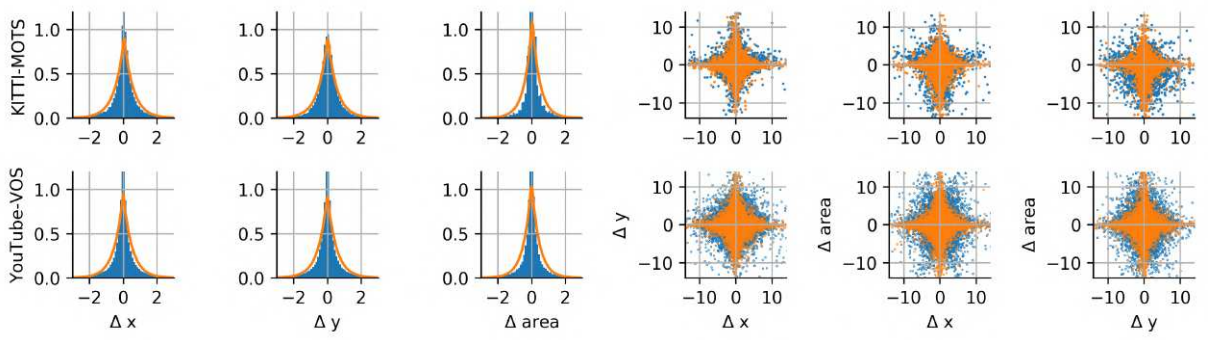


Figure 1.3: **Statistics of Natural Transitions**. Left) Distribution over transitions for horizontal (Δx) and vertical (Δy) position as well as object size ($\Delta area$) for both YouTube-VOS (Xu et al., 2018; Yang et al., 2019) and KITTI-MOTS (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016). Orange lines indicate fits of generalized Laplace distributions (Subbotin, 1923). Right) 2D marginal distribution over pairs of factor transitions (blue) and permuted pairs (orange) that indicate the marginal distributions when made independent. Figure from (Klindt et al., 2021) (made with David Klindt and Lukas Schott).

holds even if these assumptions are severely violated. For example, theoretically, we require that the FoVs underlying the data are sampled from a uniform distribution, however, empirically, we simply require that this distribution is less concentrated than the distribution underlying the positive pair used for maximizing alignment. Otherwise, it is impossible to distinguish negative pairs from positive pairs in contrastive learning.

While we empirically find that the conditions needed for successful disentanglement are significantly looser than the theoretical assumptions, we do not observe disentanglement in practice from state-of-the-art approaches due to the chosen data augmentations. Disentanglement was observed in a synthetic setting, where positive pairs were not data augmentations of the same image, but instead, image renderings corresponding to FoVs sampled from specified probability distributions, i.e. if $(\mathbf{z}, \tilde{\mathbf{z}})$ underly the positive pair, $\tilde{\mathbf{z}}$ is sampled from a distribution, e.g. truncated normal, conditioned on $\mathbf{z}$. Notably, the conditional distributions allowed all FoVs to change according to the same distribution, while data augmentations used in practice instead leave certain FoVs unchanged, leading to state-of-the-art representations preserving said FoVs, discarding the other ones. Thus, in order for visual SSL approaches to disentangle the FoVs, we must maximize alignment between positive pairs where all FoVs are able to change. Is this possible in practice?

1.3 Is disentanglement possible in practice?

In order to achieve disentanglement in practice, we must be able to fulfill the empirical conditions for success. For state-of-the-art visual SSL approaches, we require positive pairs where all FoVs can change, however, in practice, we cannot inspect the FoVs, as we only have access to the observations the FoVs underly. With that said, if we consider again the object-centric information we would like to disentangle, we recognize that the FoVs of interest can be annotated. Thus, if we can find a data source where the conditional distribution shows the FoVs are variant across positive pairs, we have identified a path to disentanglement in practice.

In our third contribution (C3) (Klindt et al., 2021), we find that *video* provides a path to

disentanglement in practice. Considering position and scale as FoVs of interest, we compute values for these FoVs from object mask annotations provided by the YouTube-VOS (Xu et al., 2018; Yang et al., 2019) and KITTI-MOTS (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016) video datasets. As can be seen in Figure 1.3 (left), if we plot the distribution for the temporal difference between the FoVs at time t and time $t-1$, we find that not only do all FoVs change between frames, but for all distributions, the temporal change is sparse. In line with the theoretical conditions presented in (Zimmermann et al., 2021), if we maximize alignment between frames t and $t-1$ using an L_1 metric, which corresponds to the conditional distribution over the representations of the positive pair that matches the observed sparse transitions of the FoVs, we theoretically and empirically achieve disentanglement.

However, real-world disentanglement has not yet been demonstrated. Theoretically, while evidence has been given that position and scale change sparsely over time, our model assumes all FoVs change sparsely over time, not only according to the same distribution, but independently as well. In Figure 1.3 (right), we can see that FoVs are statistically dependent, the magnitude of frame-to-frame change in position is correlated with that of scale, but empirically, we do not find that disentanglement is reliant on this statistical independence assumption. Still, while we observe good disentanglement, reaching 90.6% performance on disentangling position and scale after training on natural transition statistics, we only demonstrate disentanglement on object mask images. While state-of-the-art visual SSL scales well to natural visual complexity, SSL demonstrations which show that leveraging video improves downstream performance, e.g. in object recognition and localization, are still in their infancy (Tschannen et al., 2020; Sharma et al., 2022; Parthasarathy et al., 2023). With that said, our contribution is significant in bridging the gap between disentanglement and SSL in practice. After sharing the publications for C1-C3 (Von Kügelgen et al., 2021; Zimmermann et al., 2021; Klindt et al., 2021) we will conclude with a discussion on how the promise of disentanglement could be fulfilled.

Publications 2

2.1 Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

Julius von Kügelgen[*1,2] **Yash Sharma**[*3,4] **Luigi Gresele**[*1]

Wieland Brendel [3] **Bernhard Schölkopf**[†1] **Michel Besserve**[†1] **Francesco Locatello**[†5]

[1] Max Planck Institute for Intelligent Systems Tübingen [2] University of Cambridge

[3] Tübingen AI Center, University of Tübingen [4] IMPRS for Intelligent Systems [5] Amazon

Abstract

Self-supervised representation learning has shown remarkable success in a number of domains. A common practice is to perform data augmentation via hand-crafted transformations intended to leave the semantics of the data invariant. We seek to understand the empirical success of this approach from a theoretical perspective. We formulate the augmentation process as a latent variable model by postulating a partition of the latent representation into a *content* component, which is assumed invariant to augmentation, and a *style* component, which is allowed to change. Unlike prior work on disentanglement and independent component analysis, we allow for both nontrivial statistical and causal dependencies in the latent space. We study the identifiability of the latent representation based on pairs of views of the observations and prove sufficient conditions that allow us to identify the invariant content partition up to an invertible mapping in both generative and discriminative settings. We find numerical simulations with dependent latent variables are consistent with our theory. Lastly, we introduce *Causal3DIdent*, a dataset of high-dimensional, visually complex images with rich causal dependencies, which we use to study the effect of data augmentations performed in practice.

1 Introduction

Learning good representations of high-dimensional observations from large amounts of unlabelled data is widely recognised as an important step for more capable and data-efficient learning systems [10, 72]. Over the last decade, *self-supervised learning* (SSL) has emerged as the dominant paradigm for such unsupervised representation learning [1, 20, 21, 34, 41, 47, 48, 90, 91, 115, 122, 125, 126]. The main idea behind SSL is to extract a supervisory signal from unlabelled observations by leveraging known structure of the data, which allows for the application of supervised learning techniques. A common approach is to directly predict some part of the observation from another part (e.g., future from past, or original from corruption), thus forcing the model to learn a meaningful representation in the process. While this technique has shown remarkable success in natural language processing [13, 23, 30, 81, 84, 86, 95, 99] and speech recognition [5, 6, 100, 104], where a finite dictionary allows one to output a distribution over the missing part, such *predictive* SSL methods are not easily applied to continuous or high-dimensional domains such as vision. Here, a common approach is to learn a *joint embedding* of similar observations or *views* such that their representation is close [7, 12, 22, 44]. Different views can come, for example, from different modalities (text & speech; video & audio) or time points. As still images lack such multi-modality or temporal structure, recent advances in representation learning have relied on generating similar views by means of *data augmentation*.

[*]Joint first author. [†]Joint senior author. Correspondence to: jvk@tue.mpg.de
Code available at: https://www.github.com/ysharma1126/ssl_identifiability

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In order to be useful, data augmentation is thought to require the transformations applied to generate additional views to be generally chosen to *preserve the semantic characteristics* of an observation, while changing other "nuisance" aspects. While this intuitively makes sense and has shown remarkable empirical results, the success of data augmentation techniques in practice is still not very well understood from a theoretical perspective—despite some efforts [17, 19, 28]. In the present work, we seek to better understand the empirical success of SSL with data augmentation by formulating the generative process as a latent variable model (LVM) and studying *identifiability* of the representation, i.e., under which conditions the ground truth latent factors can provably be inferred from the data [77].

Related work and its relation to the current. Prior work on unsupervised representation learning from an LVM perspective often postulates *mutually independent latent factors*: this independence assumption is, for example, at the heart of independent component analysis (ICA) [24, 56] and disentanglement [10, 14, 18, 49, 65, 71]. Since it is impossible to identify the true latent factors without any supervisory signal in the general nonlinear case [57, 82], recent work has turned to weakly- or self-supervised approaches which leverage additional information in the form of multiple views [39, 83, 108, 129], auxiliary variables [58, 63], or temporal structure [45, 54, 55, 69]. To identify or disentangle the individual independent latent factors, it is typically assumed that there is a chance that *each factor changes* across views, environments, or time points.

Our work—being directly motivated by common practices in SSL with data augmentation—differs from these works in the following two key aspects (see Fig. 1 for an overview). First, we do not assume independence and instead *allow for both nontrivial statistical and causal relations between latent variables*. This is in line with a recently proposed [105] shift towards causal representation learning [40, 76, 85, 87, 106, 107, 112, 123, 127], motivated by the fact that many underlying variables of interest may not be independent but causally related to each other.[1] Second, instead of a scenario wherein all latent factors may change as a result of augmentation, we assume a *partition of the latent space* into two blocks: a *content* block which is shared or *invariant* across different augmented views, and a *style* block that *may change*. This is aligned with the notion that augmentations leave certain semantic aspects (i.e., content) intact and only affect style, and is thus a more appropriate assumption for studying SSL. In line with earlier work [39, 54, 57, 58, 63, 69, 82, 83, 129], we focus on the setting of continuous ground-truth latents, though we believe our results to hold more broadly.

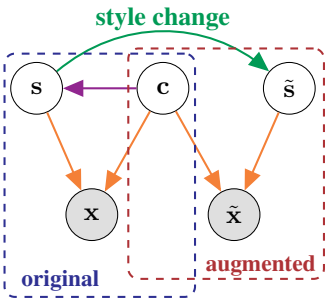


Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$.

Structure and contributions. Following a review of SSL with data augmentation and identifiability theory (§ 2), we formalise the process of data generation and augmentation as an LVM with content and style variables (§ 3). We then establish identifiability results of the invariant content partition (§ 4), validate our theoretical insights experimentally (§ 5), and discuss our findings and their limitations in the broader context of SSL with data augmentation (§ 6). We highlight the following contributions:

- we prove that SSL with data augmentations identifies the invariant content partition of the representation in generative (Thm. 4.2) and discriminative learning with invertible (Thm. 4.3) and non-invertible encoders with entropy regularisation (Thm. 4.4); in particular, Thm. 4.4 provides a theoretical justification for the empirically observed effectiveness of contrastive SSL methods that use data augmentation and InfoNCE [91] as an objective, such as `SimCLR` [20];
- we show that our theory is consistent with results in simulating statistical dependencies within blocks of content and style variables, as well as with style causally dependent on content (§ 5.1);
- we introduce *Causal3DIdent*, a dataset of 3D objects which allows for the study of identifiability in a causal representation learning setting, and use it to perform a systematic study of data augmentations used in practice, yielding novel insights on what particular data augmentations are truly isolating as invariant content and discarding as varying style when applied (§ 5.2).

[1] E.g., [69], Fig. 11 where dependence between latents was demonstrated for multiple natural video data sets.

2 Preliminaries and background

Self-supervised representation learning with data augmentation. Given an unlabelled dataset of observations (e.g., images) $\mathbf{x}$, data augmentation techniques proceed as follows. First, a set of observation-level transformations $\mathbf{t} \in \mathcal{T}$ are specified together with a distribution $p_{\mathbf{t}}$ over $\mathcal{T}$. Both $\mathcal{T}$ and $p_{\mathbf{t}}$ are typically designed using human intelligence and domain knowledge with the intention of *not changing the semantic characteristics* of the data (which arguably constitutes a form of weak supervision).[2] For images, for example, a common choice for $\mathcal{T}$ are combinations of random crops [113], horizontal or vertical flips, blurring, colour distortion [52, 113], or cutouts [31]; and $p_{\mathbf{t}}$ is a distribution over the parameterisation of these transformations, e.g., the centre and size of a crop [20, 31]. For each observation $\mathbf{x}$, a pair of transformations $\mathbf{t}, \mathbf{t}' \sim p_{\mathbf{t}}$ is sampled and applied separately to $\mathbf{x}$ to generate a pair of augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\mathbf{t}(\mathbf{x}), \mathbf{t}'(\mathbf{x}))$.

The joint-embedding approach to SSL then uses a pair of encoder functions $(\mathbf{g}, \mathbf{g}')$, i.e. deep nets, to map the pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ to a typically lower-dimensional representation $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}') = (\mathbf{g}(\tilde{\mathbf{x}}), \mathbf{g}'(\tilde{\mathbf{x}}'))$. Often, the two encoders are either identical, $\mathbf{g} = \mathbf{g}'$, or directly related (e.g., via shared parameters or asynchronous updates). Then, the encoder(s) $(\mathbf{g}, \mathbf{g}')$ are trained such that the representations $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ are "close", i.e., such that $\text{sim}(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ is large for some similarity metric $\text{sim}(\cdot)$, e.g., the cosine similarity [20, 129], or negative L2 norm [129]. The advantage of directly optimising for similarity in representation space over generative alternatives is that reconstruction can be very challenging for high-dimensional data. The disadvantage is the problem of *collapsed representations*.[3] To avoid collapsed representations and force the encoder(s) to learn a meaningful representation, two main families of approaches have been used: (i) *contrastive learning* (CL) [20, 47, 48, 91, 115, 126]; and (ii) *regularisation-based* SSL [21, 41, 128].

The idea behind CL is to not only learn similar representations for augmented views $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_i)$ of the same $\mathbf{x}_i$, or *positive pairs*, but to also use other observations $\mathbf{x}_j$ $(j \neq i)$ to contrast with, i.e., to enforce a dissimilar representation across *negative pairs* $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_j)$. In other words, CL pulls representations of positive pairs together, and pushes those of negative pairs apart. Since both aims cannot be achieved simultaneously with a constant representation, collapse is avoided. A popular CL objective function (used, e.g., in `SimCLR` [20]) is InfoNCE [91] (based on noise-contrastive estimation [42, 43]):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^{K} \sim p_{\mathbf{x}}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \right] \tag{1}$$

where $\tilde{\mathbf{z}} = \mathbb{E}_{\mathbf{t} \sim p_{\mathbf{t}}}[\mathbf{g}(\mathbf{t}(\mathbf{x}))]$, τ is a temperature, and $K-1$ is the number of negative pairs. InfoNCE (1) has an interpretation as multi-class logistic regression, and lower bounds the mutual information across similar views $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$—a common representation learning objective [4, 9, 15, 50, 75, 79, 80, 97, 120]. Moreover, (1) can be interpreted as *alignment* (numerator) and *uniformity* (denominator) terms, the latter constituting a nonparametric entropy estimator of the representation as $K \to \infty$ [124]. CL with InfoNCE can thus be seen as alignment of positive pairs with (approximate) entropy regularisation.

Instead of using negative pairs, as in CL, a set of recent SSL methods only optimise for alignment and avoid collapsed representations through different forms of regularisation. For example, `BYOL` [41] and `SimSiam` [21] rely on "architectural regularisation" in the form of moving-average updates for a separate "target" net $\mathbf{g}'$ (`BYOL` only) or a stop-gradient operation (both). `BarlowTwins` [128], on the other hand, optimises the cross correlation between $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ to be close to the identity matrix, thus enforcing redundancy reduction (zero off-diagonals) in addition to alignment (ones on the diagonal).

Identifiability of learned representations. In this work, we address the question of whether SSL with data augmentation can reveal or uncover properties of the underlying data generating process. Whether a representation learned from observations can be expected to match the true underlying latent factors—up to acceptable ambiguities and subject to suitable assumptions on the generative process and inference model—is captured by the notion of identifiability [77].

Within representation learning, identifiability has mainly been studied in the framework of (nonlinear) ICA which assumes a model of the form $\mathbf{x} = \mathbf{f}(\mathbf{z})$ and aims to recover the independent latents, or *sources*, $\mathbf{z}$, typically up to permutation or element-wise transformation. A crucial negative result states that, with i.i.d. data and without further assumptions, this is fundamentally impossible [57]. However, recent breakthroughs have shown that identifiability can be achieved if an auxiliary variable (e.g.,

[2] Note that recent work has investigated automatically discovering good augmentations [26, 27].

[3] If the only goal is to make representations of augmented views similar, a degenerate solution which simply maps any observation to the origin trivially achieves this goal.

a time stamp or environment index) renders the sources *conditionally* independent [45, 54, 55, 58]. These methods rely on constructing positive and negative pairs using the auxiliary variable and learning a representation with CL. This development has sparked a renewed interest in identifiability in the context of deep representation learning [63, 64, 69, 83, 102, 108, 109, 129].

Most closely related to SSL with data augmentation are works which study identifiability when given a second view $\tilde{\mathbf{x}}$ of an observation $\mathbf{x}$, resulting from a modified version $\tilde{\mathbf{z}}$ of the underlying latents or sources $\mathbf{z}$ [39, 69, 83, 101, 108, 129]. Here, $\tilde{\mathbf{z}}$ is either an element-wise corruption of $\mathbf{z}$ [39, 69, 101, 129] or may share a random subset of its components [83, 108]. Crucially, all previously mentioned works assume that *any* of the independent latents (are allowed to) change, and aim to identify the individual factors. However, in the context of SSL with data augmentation, where the semantic (content) part of the representation is intended to be shared between views, this assumption does not hold.

3 Problem formulation

We specify our problem setting by formalising the processes of data generation and augmentation. We take a latent-variable model perspective and assume that observations $\mathbf{x}$ (e.g., images) are generated by a *mixing* function $\mathbf{f}$ which takes a latent code $\mathbf{z}$ as input. Importantly, we describe the augmentation process through changes in this latent space as captured by a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$, as opposed to traditionally describing the transformations $\mathbf{t}$ as acting directly at the observation level.

Formally, let $\mathbf{z}$ be a continuous r.v. taking values in an open, simply-connected n-dim. *representation space* $\mathcal{Z} \subseteq \mathbb{R}^n$ with associated probability density $p_{\mathbf{z}}$. Moreover, let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a *smooth and invertible* mapping to an *observation space* $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathbf{x}$ be the continuous r.v. defined as $\mathbf{x} = \mathbf{f}(\mathbf{z})$.[4] The generative process for the dataset of original observations of $\mathbf{x}$ is thus given by:

$$\mathbf{z} \sim p_{\mathbf{z}}, \qquad \mathbf{x} = \mathbf{f}(\mathbf{z}). \tag{2}$$

Next, we formalise the data augmentation process. As stated above, we take a representation-centric view, i.e., we assume that an augmentation $\tilde{\mathbf{x}}$ of the original $\mathbf{x}$ is obtained by applying the same mixing or rendering function $\mathbf{f}$ to a modified representation $\tilde{\mathbf{z}}$ which is (stochastically) related to the original representation $\mathbf{z}$ of $\mathbf{x}$. Specifying the effect of data augmentation thus corresponds to specifying a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ which captures the relation between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

In terms of the transformation-centric view presented in § 2, we can view the modified representation $\tilde{\mathbf{z}} \in \mathcal{Z}$ as obtained by applying $\mathbf{f}^{-1}$ to a transformed observation $\tilde{\mathbf{x}} = \mathbf{t}(\mathbf{x}) \in \mathcal{X}$ where $\mathbf{t} \sim p_{\mathbf{t}}$, i.e., $\tilde{\mathbf{z}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})$. The conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ in the representation space can thus be viewed as being induced by the distribution $p_{\mathbf{t}}$ over transformations applied at the observation level.[5]

We now encode the notion that the set of transformations $\mathcal{T}$ used for augmentation is typically chosen such that any transformation $\mathbf{t} \in \mathcal{T}$ leaves certain aspects of the data invariant. To this end, we assume that *the representation* $\mathbf{z}$ *can be uniquely partitioned into two disjoint parts*:

(i) an *invariant* part $\mathbf{c}$ which will *always be shared* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *content*;
(ii) a *varying* part $\mathbf{s}$ which *may change* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *style*.

We assume that $\mathbf{c}$ and $\mathbf{s}$ take values in content and style subspaces $\mathcal{C} \subseteq \mathbb{R}^{n_c}$ and $\mathcal{S} \subseteq \mathbb{R}^{n_s}$, respectively, i.e., $n = n_c + n_s$ and $\mathcal{Z} = \mathcal{C} \times \mathcal{S}$. W.l.o.g., we let $\mathbf{c}$ corresponds to the first n_c dimensions of $\mathbf{z}$:

$$\mathbf{z} = (\mathbf{c}, \mathbf{s}), \qquad \mathbf{c} := \mathbf{z}_{1:n_c}, \qquad \mathbf{s} := \mathbf{z}_{(n_c+1):n},$$

We formalise the process of data augmentation with content-preserving transformations by defining the conditional $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ such that only a (random) subset of the style variables change at a time.

Assumption 3.1 (Content-invariance)**.** The conditional density $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ over $\mathcal{Z} \times \mathcal{Z}$ takes the form

$$p_{\tilde{\mathbf{z}}|\mathbf{z}}(\tilde{\mathbf{z}}|\mathbf{z}) = \delta(\tilde{\mathbf{c}} - \mathbf{c}) p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s})$$

for some continuous density $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ on $\mathcal{S} \times \mathcal{S}$, where $\delta(\cdot)$ is the Dirac delta function, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ a.e.

Assumption 3.2 (Style changes)**.** Let $\mathcal{A}$ be the set of subsets of style variables $A \subseteq \{1, ..., n_s\}$ and let p_A be a distribution on $\mathcal{A}$. Then, the style conditional $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ is obtained via

$$A \sim p_A, \qquad p_{\tilde{\mathbf{s}}|\mathbf{s},A}(\tilde{\mathbf{s}}|\mathbf{s}, A) = \delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A)\,,$$

where $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a continuous density on $\mathcal{S}_A \times \mathcal{S}_A$, $\mathcal{S}_A \subseteq \mathcal{S}$ denotes the subspace of changing style variables specified by A, and $A^c = \{1, ..., n_s\} \setminus A$ denotes the complement of A.

[4] While $\mathbf{x}$ may be high-dimensional $n \ll d$, invertibility of $\mathbf{f}$ implies that $\mathcal{X}$ is an n-dim. sub-manifold of $\mathbb{R}^d$.
[5] We investigate this correspondence between changes in observation and latent space empirically in § 5.

Note that Assumption 3.2 is less restrictive than assuming that all style variables need to change, since it also allows for only a (possibly different) subset of style variables to change for any given observation. This is in line with the intuition that not all transformations affect all changeable (i.e., style) properties of the data: e.g., a colour distortion should not affect positional information, and, in the same vein, a (horizontal or vertical) flip should not affect the colour spectrum.

The generative process of an augmentation or transformed observation $\tilde{\mathbf{x}}$ is thus given by

$$A \sim p_A, \qquad \tilde{\mathbf{z}}|\mathbf{z}, A \sim p_{\tilde{\mathbf{z}}|\mathbf{z},A}, \qquad \tilde{\mathbf{x}} = \mathbf{f}(\tilde{\mathbf{z}}). \tag{3}$$

Our setting for modelling data augmentation differs from that commonly assumed in (multi-view) disentanglement and ICA in that *we do not assume that the latent factors* $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ *are mutually (or conditionally) independent*, i.e., we allow for *arbitrary* (non-factorised) marginals $p_{\mathbf{z}}$ in (2).[6]

Causal interpretation: data augmentation as counterfactuals under soft style intervention. We now provide a causal account of the above data generating process by describing the (allowed) causal dependencies among latent variables using a structural causal model (SCM) [94]. As we will see, this leads to an interpretation of data augmentations as counterfactuals in the underlying latent SCM. The assumption that $\mathbf{c}$ stays invariant as $\mathbf{s}$ changes is consistent with the view that content may causally influence style, $\mathbf{c} \to \mathbf{s}$, but not vice versa, see Fig. 1. We therefore formalise their relation as:

$$\mathbf{c} := \mathbf{f}_{\mathbf{c}}(\mathbf{u}_{\mathbf{c}}), \qquad \mathbf{s} := \mathbf{f}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}), \qquad (\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}) \sim p_{\mathbf{u}_{\mathbf{c}}} \times p_{\mathbf{u}_{\mathbf{s}}}$$

where $\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}$ are independent exogenous variables, and $\mathbf{f}_{\mathbf{c}}, \mathbf{f}_{\mathbf{s}}$ are deterministic functions. The latent causal variables $(\mathbf{c}, \mathbf{s})$ are subsequently decoded into observations $\mathbf{x} = \mathbf{f}(\mathbf{c}, \mathbf{s})$. Given a factual observation $\mathbf{x}^{\mathtt{F}} = \mathbf{f}(\mathbf{c}^{\mathtt{F}}, \mathbf{s}^{\mathtt{F}})$ which resulted from $(\mathbf{u}_{\mathbf{c}}^{\mathtt{F}}, \mathbf{u}_{\mathbf{s}}^{\mathtt{F}})$, we may ask the counterfactual question: "*what would have happened if the style variables had been (randomly) perturbed, all else being equal?*". Consider, e.g., a *soft intervention* [35] on $\mathbf{s}$, i.e., an intervention that changes the mechanism $\mathbf{f}_{\mathbf{s}}$ to

$$do(\mathbf{s} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}, \mathbf{u}_A)),$$

where $\mathbf{u}_A$ is an additional source of stochasticity accounting for the randomness of the augmentation process ($p_A \times p_{\tilde{\mathbf{s}}|\mathbf{s},A}$). The resulting distribution over counterfactual observations $\mathbf{x}^{\mathtt{CF}} = \mathbf{f}(\mathbf{c}^{\mathtt{F}}, \mathbf{s}^{\mathtt{CF}})$ can be computed from the modified SCM by fixing the exogenous variables to their factual values and performing the soft intervention:

$$\mathbf{c}^{\mathtt{CF}} := \mathbf{c}^{\mathtt{F}}, \qquad \mathbf{s}^{\mathtt{CF}} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}^{\mathtt{F}}, \mathbf{u}_{\mathbf{s}}^{\mathtt{F}}, \mathbf{u}_A), \qquad \mathbf{u}_A \sim p_{\mathbf{u}_A}.$$

This aligns with our intuition and assumed problem setting of data augmentations as style corruptions. We note that the notion of augmentation as (hard) style interventions is also at the heart of `ReLIC` [87], a recently proposed, causally-inspired SSL regularisation term for instance-discrimination [44, 126]. However, `ReLIC` assumes independence between content and style and does not address identifiability. For another causal perspective on data augmentation in the context of domain generalisation, c.f. [59].

4 Theory: block-identifiability of the invariant content partition

Our goal is to prove that we can identify the invariant content partition $\mathbf{c}$ under a distinct, weaker set of assumptions, compared to existing results in disentanglement and nonlinear ICA [39, 69, 83, 108, 129]. We stress again that our primary interest is not to identify or disentangle individual (and independent) latent factors z_j, but instead to separate content from style, such that the content variables can be subsequently used for downstream tasks. We first define this distinct notion of *block-identifiability*.

Definition 4.1 (Block-identifiability)**.** We say that the true content partition $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ is *block-identified* by a function $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ if the inferred content partition $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})_{1:n_c}$ contains *all* and *only* information about $\mathbf{c}$, i.e., if there exists an *invertible* function $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

Defn. 4.1 is related to independent subspace analysis [16, 53, 73, 114], which also aims to identify blocks of random variables as opposed to individual factors, though under an *independence assumption across blocks*, and typically not within a multi-view setting as studied in the present work.

4.1 Generative self-supervised representation learning

First, we consider *generative* SSL, i.e., fitting a generative model to pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented views.[7] We show that under our specified data generation and augmentation process (§ 3),

[6] The recently proposed Independently *Modulated* Component Analysis (IMCA) [64] extension of ICA is a notable exception, but only allows for trivial dependencies across $\mathbf{z}$ in the form of a shared base measure.

[7] For notational simplicity, we present our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ rather than for two augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$, as typically used in practice but it also holds for the latter, see § 6 for further discussion.

as well as suitable additional assumptions (stated and discussed in more detail below), it is possible to isolate (i.e., block-identify) the invariant content partition. Full proofs are included in Appendix A.

Theorem 4.2 (Identifying content with a generative model)**.** *Consider the data generating process described in § 3, i.e., the pairs* $(\mathbf{x}, \tilde{\mathbf{x}})$ *of original and augmented views are generated according to* (2) *and* (3) *with* $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ *as defined in Assumptions 3.1 and 3.2. Assume further that*

(i) $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *is smooth and invertible with smooth inverse (i.e., a diffeomorphism);*

(ii) $p_{\mathbf{z}}$ *is a smooth, continuous density on* $\mathcal{Z}$ *with* $p_{\mathbf{z}}(\mathbf{z}) > 0$ *almost everywhere;*

(iii) for any $l \in \{1, ..., n_s\}$, $\exists A \subseteq \{1, ..., n_s\}$ *s.t.* $l \in A$; $p_A(A) > 0$; $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ *is smooth w.r.t. both* $\mathbf{s}_A$ *and* $\tilde{\mathbf{s}}_A$*; and for any* $\mathbf{s}_A$, $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$ *in some open, non-empty subset containing* $\mathbf{s}_A$.

If, for a given n_s $(1 \leq n_s < n)$*, a generative model* $(\hat{p}_{\mathbf{z}}, \hat{p}_A, \hat{p}_{\tilde{\mathbf{s}}|\mathbf{s},A}, \hat{\mathbf{f}})$ *assumes the same generative process (§ 3), satisfies the above assumptions (i)-(iii), and matches the data likelihood,*

$$p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) = \hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) \qquad \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X},$$

then it block-identifies the true content variables via $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ *in the sense of Defn. 4.1.*

Proof sketch. First, show (using *(i)* and the matching likelihoods) that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by $\mathbf{g}$ is related to the true $\mathbf{z}$ by a smooth invertible mapping $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ such that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c}$ is invariant across $(\mathbf{z}, \tilde{\mathbf{z}})$ almost surely w.r.t. $p_{\mathbf{z},\tilde{\mathbf{z}}}$.[8] Second, show by contradiction (using *(ii)*, *(iii)*) that $\mathbf{h}(\cdot)_{1:n_c}$ can, in fact, only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$, for otherwise the invariance from step 1 would be violated in a region of the style (sub)space of measure greater than zero.

Intuition. Thm. 4.2 assumes that the number of content (n_c) and style (n_s) variables is known, and that there is a positive probability that each *style* variable may change, though not necessarily on its own, according to *(iii)*. In this case, training a generative model of the form specified in § 3 (i.e., with an invariant content partition and subsets of changing style variables) by maximum likelihood on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ will asymptotically (in the limit of infinite data) recover the true invariant content partition up to an invertible function, i.e., it isolates, or unmixes, content from style.

Discussion. The identifiability result of Thm. 4.2 for generative SSL is of potential relevance for existing variational autoencoder (VAE) [68] variants such as the `GroupVAE` [51],[9] or its adaptive version `AdaGVAE` [83]. Since, contrary to existing results, Thm. 4.2 does not assume independent latents, it may also provide a principled basis for generative causal representation learning algorithms [76, 107, 127]. However, an important limitation to its practical applicability is that generative modelling does not tend to scale very well to complex high-dimensional observations, such as images.

4.2 Discriminative self-supervised representation learning

We therefore next turn to a discriminative approach, i.e., directly learning an encoder function $\mathbf{g}$ which leads to a similar embedding across $(\mathbf{x}, \tilde{\mathbf{x}})$. As discussed in § 2, this is much more common for SSL with data augmentations. First, we show that if an invertible encoder $\mathbf{g}$ is used, then learning a representation which is aligned in the first n_c dimensions is sufficient to block-identify content.

Theorem 4.3 (Identifying content with an invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *be any smooth and* invertible *function which minimises the following functional:*

$$\mathcal{L}_{\text{Align}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left|\left| \mathbf{g}(\mathbf{x})_{1:n_c} - \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \right|\right|_2^2 \right] \tag{4}$$

Then $\mathbf{g}$ *block-identifies the true content variables in the sense of Definition 4.1.*

Proof sketch. First, we show that the global minimum of (4) is reached by the smooth invertible function $\mathbf{f}^{-1}$. Thus, any other minimiser $\mathbf{g}$ must satisfy the same invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ used in step 1 of the proof of Thm. 4.2. The second step uses the same argument by contradiction as in Thm. 4.2.

Intuition. Thm. 4.3 states that if—under the same assumptions on the generative process as in Thm. 4.2—we directly learn a representation with an *invertible* encoder, then enforcing alignment between the first n_c latents is sufficient to isolate the invariant content partition. Intuitively, invertibility guarantees that all information is preserved, thus avoiding a collapsed representation.

[8] This step is partially inspired by [83]; the technique used to prove the second *main* step is entirely novel.

[9] which also uses a fixed content-style partition for multi-view data, but assumes that all latent factors are mutually independent, and that all style variables change between views, independent of the original style;

Discussion. According to Thm. 4.3, content can be isolated if, e.g., a flow-based architecture [32, 33, 67, 92, 93] is used, or invertibility is enforced otherwise during training [8, 60]. However, the applicability of this approach is limited as it *places strong constraints on the encoder architecture which makes it hard to scale these methods up to high-dimensional settings*. As discussed in § 2, state-of-the-art SSL methods such as `SimCLR` [20], `BYOL` [41], `SimSiam` [21], or `BarlowTwins` [128] do not use invertible encoders, but instead avoid collapsed representations—which would result from naively optimising (4) for arbitrary, non-invertible $\mathbf{g}$—using different forms of regularisation.

To close this gap between theory and practice, finally, we investigate how to block-identify content without assuming an invertible encoder. We show that, if we add a regularisation term to (4) that encourages maximum entropy of the learnt representation, the invertibility assumption can be dropped.

Theorem 4.4 (Identifying content with discriminative learning and a non-invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be any smooth function which minimises the following functional:*

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left|\left| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right|\right|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) \tag{5}$$

where $H(\cdot)$ denotes the differential entropy of the random variable $\mathbf{g}(\mathbf{x})$ taking values in $(0,1)^{n_c}$. Then $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 4.1.

Proof sketch. First, use the Darmois construction [29, 57] to build a function $\mathbf{d} : \mathcal{C} \to (0,1)^{n_c}$ mapping $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ to a uniform random variable. Then $\mathbf{g}^\star = \mathbf{d} \circ \mathbf{f}^{-1}_{1:n_c}$ attains the global minimum of (5) because $\mathbf{c}$ is invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$ and the uniform distribution is the maximum entropy distribution on $(0,1)^{n_c}$. Thus, any other minimiser $\mathbf{g}$ of (5) must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ and map to a uniform r.v. Then, use the same step 2 as in Thms. 4.2 and 4.3 to show that $\mathbf{h} = \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to (0,1)^{n_c}$ cannot depend on style, i.e., it is a function from $\mathcal{C}$ to $(0,1)^{n_c}$. Finally, we show that $\mathbf{h}$ must be invertible since it maps $p_{\mathbf{c}}$ to a uniform distribution, using a result from [129].

Intuition. Thm. 4.4 states that if we do not explicitly enforce invertibility of $\mathbf{g}$ as in Thm. 4.3, additionally maximising the entropy of the learnt representation (i.e., optimising alignment *and* uniformity [124]) avoids a collapsed representation and recovers the invariant content block. Intuitively, this is because any function that only depends on $\mathbf{c}$ will be invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$, so it is beneficial to preserve all content information to maximise entropy.

Discussion. Of our theoretical results, Thm. 4.4 requires the weakest set of assumptions, and is most closely aligned with common SSL practice. As discussed in § 2, contrastive SSL with negative samples using InfoNCE (1) as an objective can asymptotically be understood as alignment with entropy regularisation [124], i.e., objective (5). *Thm. 4.4 thus provides a theoretical justification for the empirically observed effectiveness of CL with InfoNCE*: subject to our assumptions, CL with InfoNCE asymptotically isolates content, i.e., the part of the representation that is always left invariant by augmentation. For example, the strong image classification performance based on representations learned by `SimCLR` [20], which uses color distortion and random crops as augmentations, can be explained in that object class is a content variable in this case. We extensively evaluate the effect of various augmentation techniques on different ground-truth latent factors in our experiments in § 5. There is also an interesting connection between Thm. 4.4 and `BarlowTwins` [128], which only uses positive pairs and combines alignment with a redundancy reduction regulariser that enforces decorrelation between the inferred latents. Intuitively, redundancy reduction is related to increased entropy: $\mathbf{g}^\star$ constructed in the proof of Thm. 4.4—and thus also any other minimiser of (5)—attains the global optimum of the `BarlowTwins` objective, though the reverse implication may not hold.

5 Experiments

We perform two main experiments. First, we numerically test our main result, Thm. 4.4, in a *fully-controlled*, finite sample setting (§ 5.1), using CL to estimate the entropy term in (5). Second, we seek to better understand the effect of data augmentations used *in practice* (§ 5.2). To this end, we introduce a new dataset of 3D objects with dependencies between a number of known ground-truth factors, and use it to evaluate the effect of different augmentation techniques on what is identified as content. Additional experiments are summarised in § 5.3 and described in more detail in Appendix C.

5.1 Numerical data

Experimental setup. We generate synthetic data as described in § 3. We consider $n_c = n_s = 5$, with content and style latents distributed as $\mathbf{c} \sim \mathcal{N}(0, \Sigma_c)$ and $\mathbf{s}|\mathbf{c} \sim \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_s)$, thus allowing

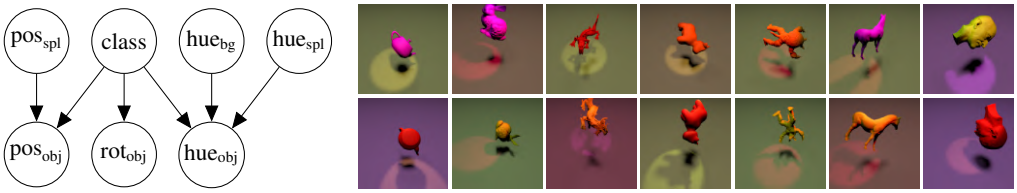


Figure 2: *(Left)* Causal graph for the *Causal3DIdent* dataset. *(Right)* Two samples from each object class.

for *statistical dependence* within the two blocks (via Σ_c and Σ_s) and *causal dependence* between content and style (via B). For $\mathbf{f}$, we use a 3-layer MLP with LeakyReLU activation functions.[10] The distribution p_A over subsets of changing style variables is obtained by independently flipping the same biased coin for each s_i. The conditional style distribution is taken as $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A} = \mathcal{N}(\mathbf{s}_A, \Sigma_A)$. We train an encoder $\mathbf{g}$ on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ with InfoNCE using the negative L2 loss as the similarity measure, i.e., we approximate (5) using empirical averages and negative samples. For evaluation, we use kernel ridge regression [88] to predict the ground truth $\mathbf{c}$ and $\mathbf{s}$ from the learnt representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$ and report the R^2 coefficient of determination. For a more detailed account, we refer to Appendix D.

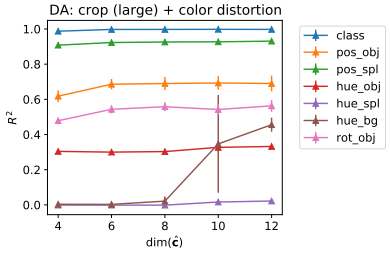
Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content $\mathbf{c}$	Style $\mathbf{s}$
1.0	✗	✗	**1.00** $\pm$ 0.00	0.07 $\pm$ 0.00
0.75	✗	✗	**1.00** $\pm$ 0.00	0.06 $\pm$ 0.05
0.75	✓	✗	**0.98** $\pm$ 0.03	0.37 $\pm$ 0.05
0.75	✓	✓	**0.99** $\pm$ 0.01	**0.80** $\pm$ 0.08

Results. In the inset table, we report mean $\pm$ std. dev. over 3 random seeds across four generative processes of increasing complexity covered by Thm. 4.4: "p(chg.)", "Stat.", and "Cau." denote respectively the change probability for each s_i, statistical dependence within blocks ($\Sigma_c \neq I \neq \Sigma_s$), and causal dependence of style on content ($B \neq 0$). An R^2 close to one indicates that almost all variation is explained by $\hat{\mathbf{c}}$, i.e., that there is a 1-1 mapping, as required by Defn. 4.1. As can be seen, *across all settings, content is block-identified*. Regarding style, we observe an increased score with the introduction of dependencies, which we explain in an extended discussion in Appendix C.1. Finally, we show in Appendix C.1 that a high R^2 score can be obtained even if we use linear regression to predict $\mathbf{c}$ from $\hat{\mathbf{c}}$ ($R^2 = 0.98 \pm 0.01$, for the last row).

5.2 High-dimensional images: *Causal3DIdent*

***Causal3DIdent* dataset.** *3DIdent* [129] is a benchmark for evaluating identifiability with rendered 224×224 images which contains hallmarks of natural environments (e.g. shadows, different lighting conditions, a 3D object). For influence of the latent factors on the renderings, see Fig. 2 of [129]. In *3DIdent*, there is a single object class (Teapot [89]), and all 10 latents are sampled independently. For *Causal3DIdent*, we introduce **six** additional classes: Hare [121], Dragon [110], Cow [62], Armadillo [70], Horse [98], and Head [111]; and impose a causal graph over the latent variables, see Fig. 2. While object class and all environment variables (spotlight position & hue, background hue) are sampled independently, all object latents are dependent,[11] see Appendix B for details.[12]

Experimental setup. For $\mathbf{g}$, we train a convolutional encoder composed of a ResNet18 [46] and an additional fully-connected layer, with LeakyReLU activation. As in `SimCLR` [20], we use InfoNCE with cosine similarity, and train on pairs of augmented examples $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$. As n_c is unknown and variable depending on the augmentation, we fix $\dim(\hat{\mathbf{c}}) = 8$ throughout. Note that we find the results to be, for the most part, robust to the choice of $\dim(\hat{\mathbf{c}})$, see inset figure. We consider the following data augmentations (DA): crop, resize & flip; colour distortion (jitter & drop); and rotation $\in \{90°, 180°, 270°\}$. For comparison, we also consider directly imposing a content-style



[10] chosen to lead to invertibility almost surely by following the settings used by previous work [54, 55]

[11] e.g., our causal graph entails hares blend into the environment (object hue centered about background & spotlight hue), a form of active camouflage observed in Alaskan [78], Arctic [2], & Snowshoe hares.

[12] We made the Causal3DIdent dataset publicly available at this URL.

Table 1: *Causal3DIdent* results: R^2 mean $\pm$ std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated, see Appendix C for the full table.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	0.42 ± 0.01	**0.61** ± 0.10	0.17 ± 0.00	0.10 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.02
LT: change hues	**1.00** ± 0.00	**0.59** ± 0.33	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.09 ± 0.08	0.21 ± 0.13	**0.87** ± 0.00	0.09 ± 0.02	**1.00** ± 0.00	0.02 ± 0.02
DA: crop (small)	0.14 ± 0.00	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00
LT: change positions	**1.00** ± 0.00	0.16 ± 0.23	0.00 ± 0.01	0.46 ± 0.02	0.00 ± 0.00	**0.97** ± 0.00	0.29 ± 0.01
DA: crop (large) + colour distortion	**0.97** ± 0.00	**0.59** ± 0.07	**0.59** ± 0.05	0.28 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	**0.74** ± 0.03
DA: crop (small) + colour distortion	**1.00** ± 0.00	**0.69** ± 0.04	**0.93** ± 0.00	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.03	**0.56** ± 0.03
LT: change positions + hues	**1.00** ± 0.00	0.22 ± 0.22	0.07 ± 0.08	0.32 ± 0.02	0.00 ± 0.01	0.02 ± 0.03	0.34 ± 0.06
DA: rotation	0.33 ± 0.06	0.17 ± 0.09	0.23 ± 0.12	**0.83** ± 0.01	0.30 ± 0.12	**0.99** ± 0.00	0.05 ± 0.03
LT: change rotations	**1.00** ± 0.00	**0.53** ± 0.33	**0.90** ± 0.00	0.41 ± 0.00	0.00 ± 0.00	**0.97** ± 0.00	0.28 ± 0.00
DA: rotation + colour distortion	**0.59** ± 0.01	**0.58** ± 0.06	0.21 ± 0.01	0.12 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.04
LT: change rotations + hues	**1.00** ± 0.00	**0.57** ± 0.34	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.00

partition by performing a latent transformation (LT) to generate views. For evaluation, we use linear logistic regression to predict object class, and kernel ridge to predict the other latents from $\hat{\mathbf{c}}$.[13]

Results. The results are presented in Tab. 1. Overall, our main findings can be summarised as:

(i) it can be difficult to design image-level augmentations that leave *specific* latent factors invariant;

(ii) augmentations & latent transformations generally have a similar effect on groups of latents;

(iii) augmentations that yield good classification performance induce variation in all other latents.

We observe that, similar to directly varying the hue latents, colour distortion leads to a discarding of hue information as style, and a preservation of (object) position as content. Crops, similar to varying the position latents, lead to a discarding of position as style, and a preservation of background and object hue as content, the latter assuming crops are sufficiently large. In contrast, image-level rotation affects both the object rotation and position, and thus deviates from only varying the rotation latents.

Whereas class is always preserved as content when generating views with latent transformations, when using data augmentations, we can only reliably decode class when crops & colour distortion are used in conjunction—a result which mirrors evaluation on ImageNet [20]. As can be seen by our evaluation of crops & colour distortion in isolation, while colour distortion leads to a discarding of hues as style, crops lead to a discarding of position & rotation as style. Thus, when used in conjunction, class is isolated as the sole content variable. See Appendix C.2 for additional analysis.

5.3 Additional experiments and ablations

We also perform an ablation on $\dim(\hat{\mathbf{c}})$ for the synthetic setting from § 5.1, see Appendix C.1 for details. Generally, we find that if $\dim(\hat{\mathbf{c}}) < n_c$, there is insufficient capacity to encode all content, so a lower-dimensional mixture of content is learnt. Conversely, if $\dim(\hat{\mathbf{c}}) > n_c$, the excess capacity is used to encode some style information (as that increases entropy). Further, we repeat our analysis from § 5.2 using `BarlowTwins` [128] (instead of `SimCLR`) which, as discussed at the end of § 4.2, is also loosely related to Thm. 4.4. The results mostly mirror those obtained for `SimCLR` and presented in Tab. 1, see Appendix C.2 for details. Finally, we ran the same experimental setup as in § 5.2 also on the *MPI3D-real* dataset [38] containing > 1 million *real* images with ground-truth annotations of 3D objects being moved by a robotic arm. Subject to some caveats, the results show a similar trend as those on *Causal3DIdent*, see Appendix C.3 for details.

6 Discussion

Theory vs practice. We have made an effort to tailor our problem formulation (§ 3) to the setting of data augmentation with content-preserving transformations. However, some of our more technical assumptions, which are necessary to prove block-identifiability of the invariant content partition, may not hold exactly in practice. This is apparent, e.g., from our second experiment (§ 5.2), where we observe that—while class should, in principle, always be invariant across views (i.e., content)—when

[13] See Appendix C.2 for results with linear regression, as well as evaluation using a higher-dimensional intermediate layer by considering a projection head [20].

using *only* crops, colour distortion, or rotation, $\mathbf{g}$ appears to encode *shortcuts* [37, 96].[14] Data augmentation, unlike latent transformations, generates views $\tilde{\mathbf{x}}$ which are not restricted to the 11-dim. image manifold $\mathcal{X}$ corresponding to the generative process of *Causal3DIdent*, but may introduce additional variation: e.g., colour distortion leads to a rich combination of colours, typically a 3-dim. feature, whereas *Causal3DIdent* only contains one degree of freedom (hue). With additional factors, any introduced invariances may be encoded as content in place of class. Image-level augmentations also tend to change multiple latent factors in a correlated way, which may violate assumption *(iii)* of our theorems, i.e., that $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is fully-supported locally. We also assume that $\mathbf{z}$ is continuous, even though *Causal3DIdent* and most disentanglement datasets also contain discrete latents. This is a very common assumption in the related literature [39, 54, 57, 58, 63, 69, 82, 83, 129] that may be relaxed in future work. Moreover, our theory holds asymptotically and at the global optimum, whereas in practice we solve a non-convex optimisation problem with a finite sample and need to approximate the entropy term in (5), e.g., using a finite number of negative pairs. The resulting challenges for optimisation may be further accentuated by the higher dimensionality of $\mathcal{X}$ induced by image-level augmentations. Finally, we remark that while, for simplicity, we have presented our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented examples, in practice, using pairs $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ of two augmented views typically yields better performance. All of our assumptions (content invariance, changing style, etc) and theoretical results still apply to the latter case. We believe that using two augmented views helps because it leads to *increased variability* across the pair: for if $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ differ from $\mathbf{x}$ in style subsets A and A', respectively, then $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ differ from each other (a.s.) in the union $A \cup A'$.

Beyond entropy regularisation. We have shown a clear link between an identifiable maximum entropy approach to SSL (Thm. 4.4) and `SimCLR` [20] based on the analysis of [124], and have discussed an intuitive connection to the notion of redundancy reduction used in `BarlowTwins` [128]. Whether other types of regularisation such as the architectural approach pursued in `BYOL` [41] and `SimSiam` [21] can also be linked to entropy maximisation, remains an open question. Deriving similar results to Thm. 4.4 with other regularisers is a promising direction for future research, c.f. [116].

The choice of augmentation technique implicitly defines content and style. As we have defined content as the part of the representation which is always left invariant across views, the choice of augmentation implicitly determines the content-style partition. This is particularly important to keep in mind when applying SSL with data augmentation to safety-critical domains, such as medical imaging. We also advise caution when using data augmentation to identify specific latent properties, since, as observed in § 5.2, image-level transformations may affect the underlying ground-truth factors in unanticipated ways. Also note that, *for a given downstream task*, we may not want to discard all style information since style variables may still be correlated with the task of interest and may thus help improve predictive performance. *For arbitrary downstream tasks*, however, where style may change in an adversarial way, it can be shown that only using content is optimal [103].

***What* vs *how* information is encoded.** We focus on *what* information is learnt by SSL with data augmentations by specifying a generative process and studying identifiability of the latent representation. Orthogonal to this, a different line of work instead studies *how* information is encoded by analysing the sample complexity needed to solve a *given downstream task* using a *linear* predictor [3, 74, 116–119]. Provided that downstream tasks only involve content, we can draw some comparisons. Whereas our results recover content only up to arbitrary invertible nonlinear functions (see Defn. 4.1), our problem setting is more general: [3, 74] assume (approximate) independence of views $(\mathbf{x}, \tilde{\mathbf{x}})$ given the task (content), while [118, 119] assume (approximate) independence between one view and the task (content) given the other view, neither of which hold in our setting.

Conclusion. Existing representation learning approaches typically assume mutually independent latents, though dependencies clearly exist in nature [106]. We demonstrate that in a *non-i.i.d.* scenario, e.g., by constructing multiple views of the same example with data augmentation, we can learn useful representations in the presence of this neglected phenomenon. More specifically, the present work contributes, to the best of our knowledge, the first: (i) identifiability result under *arbitrary dependence* between latents; and (ii) empirical study that evaluates the effect of data augmentations not only on classification, but also on other *continuous* ground-truth latents. Unlike existing identifiability results which rely on *change* as a learning signal, our approach aims to identify what is always shared across views, i.e., also using *invariance* as a learning signal. We hope that this change in perspective will be helpful for applications such as optimal style transfer or disentangling shape from pose in vision, and inspire other types of *counterfactual training* to recover a more fine-grained causal representation.

[14] class is distinguished by shape, a feature commonly unused in downstream tasks on natural images [36]

Acknowledgements

We thank: the anonymous reviewers for several helpful suggestions that triggered improvements in theory and additional experiments; Cian Eastwood, Ilyes Khemakem, Michael Lohaus, Osama Makansi, Ricardo Pio Monti, Roland Zimmermann, Weiyang Liu, and the MPI Tübingen causality group for helpful discussions and comments; Hugo Yèche for pointing out a mistake in §2 of an earlier version of the manuscript; github user `TangTangFei` for catching a bug in the implementation of the experiments from § 5.1 (that has been corrected in this version); and the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting YS.

Funding Transparency Statement

WB acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under grant no. BR 6382/1-1 as well as support by Open Philantropy and the Good Ventures Foundation. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, 01IS18039B; and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 37–45, 2015.

[2] Arctic Wildlife. Churchill Polar Bears . https://churchillpolarbears.org/churchill/, 2021.

[3] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. In *36th International Conference on Machine Learning*, pages 9904–9923. International Machine Learning Society (IMLS), 2019.

[4] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15509–15519, 2019.

[5] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33*, 2020.

[7] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[8] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

[9] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[10] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2021.

[12] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6: 737–744, 1993.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[14] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

[15] J-F Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, 4(4):112–114, 1997.

[16] Michael A Casey and Alex Westner. Separation of mixed audio sources by independent subspace analysis. In *ICMC*, pages 154–161, 2000.

[17] Olivier Chapelle and Bernhard Schölkopf. Incorporating invariances in nonlinear SVMs. In *Advances in Neural Information Processing Systems 14*, pages 609–616, Cambridge, MA, USA, 2002. MIT Press.

[18] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Advances in Neural Information Processing Systems*, pages 2615–2625, 2018.

[19] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

[20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[21] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[22] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546, 2005.

[23] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505, 2011.

[24] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[25] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[26] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[27] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[28] Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537. PMLR, 2019.

[29] G Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[31] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[32] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[33] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations*, 2017.

[34] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[35] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74 (5):981–995, 2007.

[36] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*, 2019.

[37] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

[38] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32:15740–15751, 2019.

[39] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, 2019.

[40] Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

[41] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33*, 2020.

[42] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[43] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

[44] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006.

[45] Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[47] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[48] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 2020.

[49] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

[50] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations*, 2019.

[51] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pages 2506–2513, 2019.

[52] Andrew G. Howard. Some improvements on deep convolutional neural network based image classification, 2013.

[53] Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

[54] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

[55] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[56] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[57] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[58] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[59] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International Conference on Machine Learning*, pages 4555–4562. PMLR, 2021.

[60] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.

[61] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, 1982.

[62] Keenan's 3D Model Repository. Keenan's 3D Model Repository . https://www.cs.cmu.edu/ kmcrane/Projects/ModelRepository/, 2021.

[63] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 108, pages 2207–2217, 2020.

[64] Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems 33*, 2020.

[65] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[66] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

[67] Diederik P Kingma and Prafulla Dhariwal. Glow: generative flow with invertible $1\times$ 1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[68] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[69] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *International Conference on Learning Representations (ICLR)*, 2021.

[70] Venkat Krishnamurthy and Marc Levoy. Fitting smooth surfaces to dense polygon meshes. In John Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 313–324. ACM, 1996.

[71] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[72] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[73] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, 2011.

[74] Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

[75] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11 (2):417–441, 1999.

[76] Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *arXiv preprint arXiv:2006.07796*, 2020.

[77] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[78] Lepus Othus. Animal Diversity Web . https://animaldiversity.org/accounts/Lepus_othus/, 2021.

[79] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[80] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, pages 186–194, 1989.

[81] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[82] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

[83] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6348–6359. PMLR, 2020.

[84] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations*, 2018.

[85] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

[86] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[87] Jovana Mitrovic, Brian McWilliams, Jacob C. Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *9th International Conference on Learning Representations*, 2021.

[88] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, MA, 2012.

[89] Martin Edward Newell. *The Utilization of Procedure Models in Digital Image Synthesis.* PhD thesis, The University of Utah, 1975. AAI7529894.

[90] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[91] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[92] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[93] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[94] Judea Pearl. *Causality*. Cambridge university press, 2009.

[95] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[96] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *arXiv preprint arXiv:2011.09468*, 2020.

[97] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[98] Emil Praun, Adam Finkelstein, and Hugues Hoppe. Lapped textures. In Judith R. Brown and Kurt Akeley, editors, *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 465–470. ACM, 2000.

[99] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018.

[100] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993, 2020.

[101] H. Richard, L. Gresele, A. Hyvarinen, B. Thirion, A. Gramfort, and P. Ablin. Modeling shared responses in neuroimaging studies through multiview ica. In *Advances in Neural Information Processing Systems 33*, pages 19149–19162. Curran Associates, Inc., December 2020.

[102] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

[103] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[104] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, pages 3465–3469, 2019.

[105] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[106] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[107] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

[108] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *8th International Conference on Learning Representations*, 2020.

[109] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020.

[110] Stanford Scanning Repository. The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/, 2021.

[111] Suggestive Contour Gallery. Suggestive Contour Gallery . https://gfx.cs.princeton.edu/proj/sugcon/models/, 2021.

[112] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

[113] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[114] Fabian Theis. Towards a general independent subspace analysis. *Advances in Neural Information Processing Systems*, 19:1361–1368, 2006.

[115] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision - ECCV 2020*, volume 12356, pages 776–794. Springer, 2020.

[116] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning,*, volume 139, pages 10268–10278, 2021.

[117] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive estimation reveals topic posterior information to linear models. *arXiv preprint arXiv:2003.02234*, 2020.

[118] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[119] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *International Conference on Learning Representations*, 2020.

[120] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[121] Greg Turk and Marc Levoy. Zippered polygon meshes from range images. In Dino Schweitzer, Andrew S. Glassner, and Mike Keeler, editors, *Proceedings of the 21th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1994, Orlando, FL, USA, July 24-29, 1994*, pages 311–318. ACM, 1994.

[122] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

[123] Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. In *ICLR Workshop on "Causal Learning for Decision Making"*, 2020.

[124] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020.

[125] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[126] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3733–3742. IEEE Computer Society, 2018.

[127] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[128] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 12310–12320, 2021.

[129] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 12979–12990, 2021.

APPENDIX

Overview:

- Appendix A contains the full proofs for all theoretical results from the main paper.
- Appendix B contains additional details and plots for the *Causal3DIdent* dataset.
- Appendix C contains additional experimental results and analysis.
- Appendix D contains additional implementation details for our experiments.

A Proofs

We now present the full detailed proofs of our three theorems which were briefly sketched in the main paper. We remark that these proofs build on each other, in the sense that the (main) step 2 of the proof of Thm. 4.2 is also used in the proofs of Thms. 4.3 and 4.4.

A.1 Proof of Thm. 4.2

Theorem 4.2 (Identifying content with a generative model)**.** *Consider the data generating process described in § 3, i.e., the pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented views are generated according to (2) and (3) with $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ as defined in Assumptions 3.1 and 3.2. Assume further that*

(i) $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ *is smooth and invertible with smooth inverse (i.e., a diffeomorphism);*

(ii) $p_{\mathbf{z}}$ *is a smooth, continuous density on $\mathcal{Z}$ with $p_{\mathbf{z}}(\mathbf{z}) > 0$ almost everywhere;*

(iii) for any $l \in \{1, ..., n_s\}$, $\exists A \subseteq \{1, ..., n_s\}$ s.t. $l \in A$; $p_A(A) > 0$; $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is smooth w.r.t. both $\mathbf{s}_A$ and $\tilde{\mathbf{s}}_A$; and for any $\mathbf{s}_A$, $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$ in some open, non-empty subset containing $\mathbf{s}_A$.

If, for a given n_s ($1 \leq n_s < n$), a generative model $(\hat{p}_{\mathbf{z}}, \hat{p}_A, \hat{p}_{\tilde{\mathbf{s}}|\mathbf{s},A}, \hat{\mathbf{f}})$ assumes the same generative process (§ 3), satisfies the above assumptions (i)-(iii), and matches the data likelihood,

$$p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) = \hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) \qquad \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X},$$

then it block-identifies the true content variables via $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ in the sense of Defn. 4.1.

Proof. The proof consists of two main steps.

In the first step, we use assumption *(i)* and the matching likelihoods to show that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ is related to the true latent $\mathbf{z}$ by a smooth invertible mapping $\mathbf{h}$, and that $\hat{\mathbf{z}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ in the first n_c (content) components almost surely (a.s.) with respect to (w.r.t.) the true generative process.

In the second step, we then use assumptions *(ii)* and *(iii)* to prove (by contradiction) that $\hat{\mathbf{c}} := \hat{\mathbf{z}}_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}$ can, in fact, only depend on the true content $\mathbf{c}$ and not on the true style $\mathbf{s}$, for otherwise the invariance established in the first step would have be violated with probability greater than zero.

To provide some further intuition for the second step, the assumed generative process implies that $(\mathbf{c}, \mathbf{s}, \tilde{\mathbf{s}})|A$ is constrained to take values (a.s.) in a subspace $\mathcal{R}$ of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ of dimension $n_c + n_s + |A|$ (as opposed to dimension $n_c + 2n_s$ for $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$). In this context, assumption *(iii)* implies that $(\mathbf{c}, \mathbf{s}, \tilde{\mathbf{s}})|A$ has a density with respect to a measure on this subspace equivalent to the Lebesgue measure on $\mathbb{R}^{n_c+n_s+|A|}$. This equivalence implies, in particular, that this "subspace measure" is strictly positive: it takes strictly positive values on open sets of $\mathcal{R}$ seen as a topological subspace of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$. These open sets are defined by the induced topology: they are the intersection of the open sets of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ with $\mathcal{R}$. An open set B of V on which $p(\mathbf{c}, \mathbf{s}, \tilde{\mathbf{s}}|A) > 0$ then satisfies $P(B|A) > 0$. We look for such an open set to prove our result.

Step 1. From the assumed data generating process described in § 3—in particular, from the form of the model conditional $\hat{p}_{\tilde{\mathbf{z}}|\mathbf{z}}$ described in Assumptions 3.1 and 3.2—it follows that

$$\mathbf{g}(\mathbf{x})_{1:n_c} = \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \tag{6}$$

a.s., i.e., with probability one, w.r.t. the model distribution $\hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}$.

Due to the assumption of matching likelihoods, the invariance in (6) must also hold (a.s.) w.r.t. the true data distribution $p_{\mathbf{x},\tilde{\mathbf{x}}}$.

Next, since $\mathbf{f}, \hat{\mathbf{f}} : \mathcal{Z} \to \mathcal{X}$ are smooth and invertible functions by assumption *(i)*, there exists a smooth and invertible function $\mathbf{h} = \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to \mathcal{Z}$ such that

$$\mathbf{g} = \mathbf{h} \circ \mathbf{f}^{-1}. \tag{7}$$

Substituting (7) into (6), we obtain (a.s. w.r.t. p):

$$\hat{\mathbf{c}} := \hat{\mathbf{z}}_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c} = \mathbf{h}(\mathbf{f}^{-1}(\mathbf{x}))_{1:n_c} = \mathbf{h}(\mathbf{f}^{-1}(\tilde{\mathbf{x}}))_{1:n_c} \tag{8}$$

Substituting $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$ and $\tilde{\mathbf{z}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})$ into (8), we obtain (a.s. w.r.t. p)

$$\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c} = \mathbf{h}(\tilde{\mathbf{z}})_{1:n_c}. \tag{9}$$

It remains to show that $\mathbf{h}(\cdot)_{1:n_c}$ can only be a function of $\mathbf{c}$, i.e., does not depend on any other (style) dimension of $\mathbf{z} = (\mathbf{c}, \mathbf{s})$.

Step 2. Suppose *for a contradiction* that $\mathbf{h}_c(\mathbf{c}, \mathbf{s}) := \mathbf{h}(\mathbf{c}, \mathbf{s})_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}$ depends on some component of the style variable $\mathbf{s}$:

$$\exists l \in \{1, ..., n_s\}, (\mathbf{c}^*, \mathbf{s}^*) \in \mathcal{C} \times \mathcal{S}, \qquad \text{s.t.} \qquad \frac{\partial \mathbf{h}_c}{\partial s_l}(\mathbf{c}^*, \mathbf{s}^*) \neq 0, \tag{10}$$

that is, we assume that the partial derivative of $\mathbf{h}_c$ w.r.t. some style variable s_l is non-zero at some point $\mathbf{z}^* = (\mathbf{c}^*, \mathbf{s}^*) \in \mathcal{Z} = \mathcal{C} \times \mathcal{S}$.

Since $\mathbf{h}$ is smooth, so is $\mathbf{h}_c$. Therefore, $\mathbf{h}_c$ has continuous (first) partial derivatives.

By continuity of the partial derivative, $\frac{\partial \mathbf{h}_c}{\partial s_l}$ must be non-zero in a neighbourhood of $(\mathbf{c}^*, \mathbf{s}^*)$, i.e.,

$$\exists \eta > 0 \qquad \text{s.t.} \qquad s_l \mapsto \mathbf{h}_c\big(\mathbf{c}^*, (\mathbf{s}^*_{-l}, s_l)\big) \quad \text{is strictly monotonic on} \quad (s^*_l - \eta, s^*_l + \eta), \tag{11}$$

where $\mathbf{s}_{-l} \in \mathcal{S}_{-l}$ denotes the vector of remaining style variables except s_l.

Next, define the auxiliary function $\psi : \mathcal{C} \times \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ as follows:

$$\psi(\mathbf{c}, \mathbf{s}, \tilde{\mathbf{s}}) := |\mathbf{h}_c(\mathbf{c}, \mathbf{s}) - \mathbf{h}_c(\mathbf{c}, \tilde{\mathbf{s}})| \geq 0\,. \tag{12}$$

To obtain a contradiction to the invariance condition (9) from Step 1 under assumption (10), it remains to show that ψ from (12) is *strictly positive* with probability greater than zero (w.r.t. p).

First, the strict monotonicity from (11) implies that

$$\psi\big(\mathbf{c}^*, (\mathbf{s}^*_{-l}, s_l), (\mathbf{s}^*_{-l}, \tilde{s}_l)\big) > 0\,, \quad \forall (s_l, \tilde{s}_l) \in (s^*_l, s^*_l + \eta) \times (s^*_l - \eta, s^*_l)\,. \tag{13}$$

Note that in order to obtain the strict inequality in (13), it is important that s_l and $\tilde{s}_l$ take values in *disjoint* open subsets of the interval $(s^*_l - \eta, s^*_l + \eta)$ from (11).

Since ψ is a composition of continuous functions (absolute value of the difference of two continuous functions), ψ is continuous.

Consider the open set $\mathbb{R}_{>0}$, and recall that, under a continuous function, pre-images (or inverse images) of open sets are always *open*.

Applied to the continuous function ψ, this pre-image corresponds to an *open* set

$$\mathcal{U} \subseteq \mathcal{C} \times \mathcal{S} \times \mathcal{S} \tag{14}$$

in the domain of ψ on which ψ is strictly positive.

Moreover, due to (13):

$$\{\mathbf{c}^*\} \times \big(\{\mathbf{s}^*_{-l}\} \times (s^*_l, s^*_l + \eta)\big) \times \big(\{\mathbf{s}^*_{-l}\} \times (s^*_l - \eta, s^*_l)\big) \subset \mathcal{U}, \tag{15}$$

so $\mathcal{U}$ is *non-empty*.

Next, by assumption *(iii)*, there exists at least one subset $A \subseteq \{1, ..., n_s\}$ of changing style variables such that $l \in A$ and $p_A(A) > 0$; pick one such subset and call it A.

Then, also by assumption *(iii)*, for any $\mathbf{s}_A \in \mathcal{S}_A$, there is an open subset $\mathcal{O}(\mathbf{s}_A) \subseteq \mathcal{S}_A$ containing $\mathbf{s}_A$, such that $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A) > 0$ within $\mathcal{O}(\mathbf{s}_A)$.

Define the following space

$$\mathcal{R}_A := \{(\mathbf{s}_A, \tilde{\mathbf{s}}_A) : \mathbf{s}_A \in \mathcal{S}_A, \tilde{\mathbf{s}}_A \in \mathcal{O}(\mathbf{s}_A)\} \tag{16}$$

and, recalling that $A^c = \{1, ..., n_s\} \setminus A$ denotes the complement of A, define

$$\mathcal{R} := \mathcal{C} \times \mathcal{S}_{A^c} \times \mathcal{R}_A \tag{17}$$

which is a topological subspace of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$.

By assumptions *(ii)* and *(iii)*, $p_{\mathbf{z}}$ is smooth and fully supported, and $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\cdot|\mathbf{s}_A)$ is smooth and fully supported on $\mathcal{O}(\mathbf{s}_A)$ for any $\mathbf{s}_A \in \mathcal{S}_A$. Therefore, the measure $\mu_{(\mathbf{c},\mathbf{s}_{A^c},\mathbf{s}_A,\tilde{\mathbf{s}}_A)|A}$ has fully supported, strictly-positive density on $\mathcal{R}$ w.r.t. a strictly positive measure on $\mathcal{R}$. In other words, $p_{\mathbf{z}} \times p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is fully supported (i.e., strictly positive) on $\mathcal{R}$.

Now consider the intersection $\mathcal{U} \cap \mathcal{R}$ of the open set $\mathcal{U}$ with the topological subspace $\mathcal{R}$.

Since $\mathcal{U}$ is open, by the definition of topological subspaces, the intersection $\mathcal{U} \cap \mathcal{R} \subseteq \mathcal{R}$ is *open* in $\mathcal{R}$, (and thus has the same dimension as $\mathcal{R}$ if non-empty).

Moreover, since $\mathcal{O}(\mathbf{s}_A^*)$ is open containing $\mathbf{s}_A^*$, there exists $\eta' > 0$ such that $\{\mathbf{s}_{-l}^*\} \times (s_l^* - \eta', s_l^*) \subset \mathcal{O}(\mathbf{s}_A^*)$. Thus, for $\eta'' = \min(\eta, \eta') > 0$,

$$\{\mathbf{c}^*\} \times \{\mathbf{s}_{A^c}^*\} \times \Big(\{\mathbf{s}_{A\setminus\{l\}}^*\} \times (s_l^*, s_l^* + \eta)\Big) \times \Big(\{\mathbf{s}_{A\setminus\{l\}}^*\} \times (s_l^* - \eta'', s_l^*)\Big) \subset \mathcal{R}. \tag{18}$$

In particular, this implies that

$$\{\mathbf{c}^*\} \times \big(\{\mathbf{s}_{-l}^*\} \times (s_l^*, s_l^* + \eta)\big) \times \big(\{\mathbf{s}_{-l}^*\} \times (s_l^* - \eta'', s_l^*)\big) \subset \mathcal{R}, \tag{19}$$

Now, since $\eta'' \leq \eta$, the LHS of (19) is also in $\mathcal{U}$ according to (15), so the intersection $\mathcal{U} \cap \mathcal{R}$ is *non-empty*.

In summary, the intersection $\mathcal{U} \cap \mathcal{R} \subseteq \mathcal{R}$:

- is non-empty (since both $\mathcal{U}$ and $\mathcal{R}$ contain the LHS of (15));
- is an open subset of the topological subspace $\mathcal{R}$ of $\mathcal{C} \times \mathcal{S} \times \mathcal{S}$ (since it is the intersection of an open set, $\mathcal{U}$, with $\mathcal{R}$);
- satisfies $\psi > 0$ (since this holds for all of $\mathcal{U}$);
- is fully supported w.r.t. the generative process (since this holds for all of $\mathcal{R}$).

As a consequence,

$$\mathbb{P}\left(\psi(\mathbf{c}, \mathbf{s}, \tilde{\mathbf{s}}) > 0 | A\right) \geq \mathbb{P}(\mathcal{U} \cap \mathcal{R}) > 0, \tag{20}$$

where $\mathbb{P}$ denotes probability w.r.t. the true generative process p.

Since $p_A(A) > 0$, this is a **contradiction** to the invariance (9) from Step 1.

Hence, assumption (10) cannot hold, i.e., $\mathbf{h}_c(\mathbf{c}, \mathbf{s})$ does not depend on any style variable s_l. It is thus only a function of $\mathbf{c}$, i.e., $\hat{\mathbf{c}} = \mathbf{h}_c(\mathbf{c})$.

Finally, smoothness and invertibility of $\mathbf{h}_c : \mathcal{C} \to \mathcal{C}$ follow from smoothness and invertibility of $\mathbf{h}$, as established in Step 1.

This concludes the proof that $\hat{\mathbf{c}}$ is related to the true content $\mathbf{c}$ via a smooth invertible mapping. □

A.2 Proof of Thm. 4.3

Theorem 4.3 (Identifying content with an invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *be any smooth and* invertible *function which minimises the following functional:*

$$\mathcal{L}_{\text{Align}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left|\left| \mathbf{g}(\mathbf{x})_{1:n_c} - \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \right|\right|_2^2 \right] \tag{4}$$

Then $\mathbf{g}$ *block-identifies the true content variables in the sense of Definition 4.1.*

Proof. As in the proof of Thm. 4.2, the proof again consists of two main steps.

In the first step, we show that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by any $\mathbf{g}$ that minimises $\mathcal{L}_{\text{Align}}$ is related to the true latent $\mathbf{z}$ through a smooth invertible mapping $\mathbf{h}$, and that $\hat{\mathbf{z}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ in the first n_c (content) components almost surely (a.s.) with respect to (w.r.t.) the true generative process.

In the second step, we use the same argument by contradiction as in Step 2 of the proof of Thm. 4.2, to show that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c}$ can only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$.

Step 1. From the form of the objective (4), it is clear that $\mathcal{L}_{\text{Align}} \geq 0$ with equality if and only if $\mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c}$ for all $(\mathbf{x}, \tilde{\mathbf{x}})$ s.t. $p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) > 0$.

Moreover, it follows from the assumed generative process that the global minimum of zero is attained by the true unmixing $\mathbf{f}^{-1}$ since

$$\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c} = \tilde{\mathbf{c}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})_{1:n_c} \tag{21}$$

holds a.s. (i.e., with probability one) w.r.t. the true generative process p.

Hence, there exists at least one smooth invertible function ($\mathbf{f}^{-1}$) which attains the global minimum.

Let $\mathbf{g}$ be *any* function attaining the global minimum of $\mathcal{L}_{\text{Align}}$ of zero.

As argued above, this implies that (a.s. w.r.t. p):

$$\mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} = \mathbf{g}(\mathbf{x})_{1:n_c}. \tag{22}$$

Writing $\mathbf{g} = \mathbf{h} \circ \mathbf{f}^{-1}$, where $\mathbf{h}$ is the smooth, invertible function $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ we obtain (a.s. w.r.t. p):

$$\hat{\mathbf{c}} = \mathbf{h}(\tilde{\mathbf{z}})_{1:n_c} = \mathbf{h}(\mathbf{z})_{1:n_c}. \tag{23}$$

Note that this is the same invariance condition as (9) derived in Step 1 of the proof of Thm. 4.2.

Step 2. It remains to show that $\mathbf{h}(\mathbf{z})_{1:n_c}$ can only depend on the true content $\mathbf{c}$ and not on any of the style variables $\mathbf{s}$. To show this, we use the same Step 2 as in the proof of Thm. 4.2. □

A.3 Proof of Thm. 4.4

Theorem 4.4 (Identifying content with discriminative learning and a non-invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be any smooth function which minimises the following functional:*

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left|\left| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right|\right|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) \tag{5}$$

where $H(\cdot)$ denotes the differential entropy of the random variable $\mathbf{g}(\mathbf{x})$ taking values in $(0,1)^{n_c}$. Then $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 4.1.

Proof. The proof consists of three main steps.

In the first step, we show that the representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$ extracted by any smooth function $\mathbf{g}$ that minimises (5) is related to the true latent $\mathbf{z}$ through a smooth mapping $\mathbf{h}$; that $\hat{\mathbf{c}}$ must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ almost surely (a.s.) with respect to (w.r.t.) the true generative process p; and that $\hat{\mathbf{c}}$ must follow a uniform distribution on $(0,1)^{n_c}$.

In the second step, we use the same argument by contradiction as in Step 2 of the proof of Thm. 4.2, to show that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})$ can only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$.

Finally, in the third step, we show that $\mathbf{h}$ must be a bijection, i.e., invertible, using a result from [129].

Step 1. The global minimum of $\mathcal{L}_{\text{AlignMaxEnt}}$ is reached when the first term (alignment) is minimised (i.e., equal to zero) and the second term (entropy) is maximised.

Without additional moment constraints, the *unique* maximum entropy distribution on $(0,1)^{n_c}$ is the uniform distribution [25, 61].

First, we show that there exists a smooth function $\mathbf{g}^* : \mathcal{X} \to (0,1)^{n_c}$ which attains the global minimum of $\mathcal{L}_{\text{AlignMaxEnt}}$.

To see this, consider the function $\mathbf{f}_{1:n_c}^{-1} : \mathcal{X} \to \mathcal{C}$, i.e., the inverse of the true mixing $\mathbf{f}$, restricted to its first n_c dimensions. This exists and is smooth since $\mathbf{f}$ is smooth and invertible by assumption *(i)*. Further, we have $\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c}$ by definition.

We now build a function $\mathbf{d} : \mathcal{C} \to (0,1)^{n_c}$ which maps $\mathbf{c}$ to a uniform random variable on $(0,1)^{n_c}$ using a recursive construction known as the *Darmois construction* [29, 57].

Specifically, we define

$$d_i(\mathbf{c}) := F_i(c_i|\mathbf{c}_{1:i-1}) = \mathbb{P}(C_i \leq c_i|\mathbf{c}_{1:i-1}), \qquad i = 1, ..., n_c, \tag{24}$$

where F_i denotes the conditional cumulative distribution function (CDF) of c_i given $\mathbf{c}_{1:i-1}$.

By construction, $\mathbf{d}(\mathbf{c})$ is uniformly distributed on $(0,1)^{n_c}$ [29, 57].

Further, $\mathbf{d}$ is smooth by the assumption that $p_{\mathbf{z}}$ (and thus $p_{\mathbf{c}}$) is a smooth density.

Finally, we define

$$\mathbf{g}^* := \mathbf{d} \circ \mathbf{f}_{1:n_c}^{-1} : \mathcal{X} \to (0,1)^{n_c}, \tag{25}$$

which is a smooth function since it is a composition of two smooth functions.

Claim A.1. $\mathbf{g}^*$ as defined in (25) attains the global minimum of $\mathcal{L}_{\text{AlignMaxEnt}}$.

Proof of Claim A.1. Using $\mathbf{f}^{-1}(\mathbf{x})_{1:n_c} = \mathbf{c}$ and $\mathbf{f}^{-1}(\tilde{\mathbf{x}})_{1:n_c} = \tilde{\mathbf{c}}$, we have

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}^*) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left|\left| \mathbf{g}^*(\mathbf{x}) - \mathbf{g}^*(\tilde{\mathbf{x}}) \right|\right|_2^2 \right] - H\left(\mathbf{g}^*(\mathbf{x})\right) \tag{26}$$

$$= \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left|\left| \mathbf{d}(\mathbf{c}) - \mathbf{d}(\tilde{\mathbf{c}}) \right|\right|_2^2 \right] - H\left(\mathbf{d}(\mathbf{c})\right) \tag{27}$$

$$= 0 \tag{28}$$

where in the last step we have used the fact that $\mathbf{c} = \tilde{\mathbf{c}}$ almost surely w.r.t. to the ground truth generative process p described in § 3, so the first term is zero; and the fact that $\mathbf{d}(\mathbf{c})$ is uniformly distributed on $(0,1)^{n_c}$ and the uniform distribution on the unit hypercube has zero entropy, so the second term is also zero.

Next, let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be *any* smooth function which attains the global minimum of (5), i.e.,

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left|\left| \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right|\right|_2^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) = 0. \tag{29}$$

Define $\mathbf{h} := \mathbf{g} \circ \mathbf{f} : \mathcal{Z} \to (0,1)^{n_c}$ which is smooth because both $\mathbf{g}$ and $\mathbf{f}$ are smooth.

Writing $\mathbf{x} = \mathbf{f}(\mathbf{z})$, (29) then implies in terms of $\mathbf{h}$:

$$\mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{(\mathbf{x},\tilde{\mathbf{x}})}} \left[\left|\left| \mathbf{h}(\mathbf{z}) - \mathbf{h}(\tilde{\mathbf{z}}) \right|\right|_2^2 \right] = 0\,, \tag{30}$$

$$H\left(\mathbf{h}(\mathbf{z})\right) = 0\,. \tag{31}$$

Equation (30) implies that the same invariance condition (9) used in the proofs of Thms. 4.2 and 4.3 must hold (a.s. w.r.t. p), and (31) implies that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})$ must be uniformly distributed on $(0,1)^{n_c}$.

Step 2. Next, we show that $\mathbf{h}(\mathbf{z}) = \mathbf{h}(\mathbf{c}, \mathbf{s})$ can only depend on the true content $\mathbf{c}$ and not on any of the style variables $\mathbf{s}$. For this we use the same Step 2 as in the proofs of Thms. 4.2 and 4.3.

Step 3. Finally, we show that the mapping $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$ is invertible.

To this end, we make use of the following result from [129].

Proposition A.2 (Proposition 5 of [129])**.** *Let $\mathcal{M}, \mathcal{N}$ be simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries and $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map. Further, let the random variable $\mathbf{z} \in \mathcal{M}$ be distributed according to $\mathbf{z} \sim p(\mathbf{z})$ for a regular density function p, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}(\mathbf{z})$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

We apply this result to the simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries $\mathcal{M} = \mathcal{C}$ and $\mathcal{N} = (0,1)^{n_c}$, and the smooth (hence, differentiable) map $\mathbf{h} : \mathcal{C} \to (0,1)^{n_c}$ which maps the random variable $\mathbf{c}$ to a uniform random variable $\hat{\mathbf{c}}$ (as established in Step 1).

Since both $p_{\mathbf{c}}$ (by assumption) and the uniform distribution (the pushforward of $p_{\mathbf{c}}$ through $\mathbf{h}$) are regular densities in the sense of Prop. A.2, we conclude that $\mathbf{h}$ is a bijection, i.e., invertible.

We have shown that for any smooth $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ which minimises $\mathcal{L}_{\text{AlignMaxEnt}}$, we have that $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x}) = \mathbf{h}(\mathbf{c})$ for a smooth and invertible $\mathbf{h} : \mathcal{C} \to (0,1)^{n_c}$, i.e., $\mathbf{c}$ is block-identified by $\mathbf{g}$. □

B Additional details on the Causal3DIdent data set

Using the Blender rendering engine [11], 3DIdent [129] is a recently proposed benchmark which contains hallmarks of natural environments (e.g. shadows, different lighting conditions, a 3D object), but allows for identifiability evaluation by exposing the underlying generative factors.

Each $224 \times 224 \times 3$ image in the dataset shows a coloured 3D object which is located and rotated above a coloured ground in a 3D space. Furthermore, each scene contains a coloured spotlight which is focused on the object and located on a half-circle around the scene. The images are rendered based on a 10-dimensional latent, where: (i) three dimensions describe the XYZ position of the object, (ii) three dimensions describe the rotation of the object in Euler angles, (iii) two dimensions describe the colour (hue) of the object and the ground of the scene, respectively, and (iv) two dimensions describe the position and colour (hue) of the spotlight. For influence of the latent factors on the renderings, see Fig. 2 of [129].

B.1 Details on introduced object classes

3DIdent contained a single object class, Teapot [89]. We add **six** additional object classes: Hare [121], Dragon [110], Cow [62], Armadillo [70], Horse [98], Head [111].

B.2 Details on latent causal graph

In 3DIdent, the latents are uniformly sampled independently. We instead impose a causal graph over the variables (see Fig. 2). While object class and all environment variables (spotlight position, spotlight hue, background hue) are sampled independently, all object variables are dependent. Specifically, for spotlight position, spotlight hue, and background hue, we sample from $U(-1,1)$. We impose the dependence by varying the mean (μ) of a truncated normal distribution with standard deviation $\sigma = 0.5$, truncated to the range $[-1,1]$.

Object rotation is dependent solely on object class, see Tab. 2 for details. Object position is dependent on both object class & spotlight position, see Tab. 3. Object hue is dependent on object class, background hue, & object hue, see Tab. 4. Hares blending into their environment as a form of active camouflage has been observed in Alaskan [78], Arctic [2], & Snowshoe hares.

B.3 Dataset Visuals

We show 40 random samples from the marginal of each object class in Causal3DIdent in Figs. 3 to 9.

Table 2: Given a certain object class, the center of the truncated normal distribution from which we sample *rotation* latents varies.

object class	$\mu(\phi)$	$\mu(\theta)$	$\mu(\psi)$
Teapot	-0.35	0.35	0.35
Hare	0.35	-0.35	0.35
Dragon	0.35	0.35	-0.35
Cow	0.35	-0.35	-0.35
Armadillo	-0.35	0.35	-0.35
Horse	-0.35	-0.35	0.35
Head	-0.35	-0.35	-0.35

Table 3: Given a certain object class & spotlight position, the center of the truncated normal distribution from which we sample *xy-position* latents varies. Note the spotlight position pos_{spl} is rescaled from $[-1, 1]$ to $[-\pi/2, \pi/2]$.

object class	$\mu(x)$	$\mu(y)$	$\mu(z)$
Teapot	0	0	0
Hare	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Dragon	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Cow	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0
Armadillo	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0
Horse	$-\sin(\text{pos}_{\text{spl}})$	$-\cos(\text{pos}_{\text{spl}})$	0
Head	$\sin(\text{pos}_{\text{spl}})$	$\cos(\text{pos}_{\text{spl}})$	0

Table 4: Given a certain object class, background hue, and spotlight hue, the center of the truncated normal distribution from which we sample the *object hue* latent varies. Note that for the Hare and Dragon classes, in particular, the object either blends in or stands out from the environment.

object class	$\mu(\text{hue})$
Teapot	0
Hare	$\frac{\text{hue}_{\text{bg}}+\text{hue}_{\text{spl}}}{2}$
Dragon	$-\frac{\text{hue}_{\text{bg}}+\text{hue}_{\text{spl}}}{2}$
Cow	-0.35
Armadillo	0.7
Horse	-0.7
Head	0.35

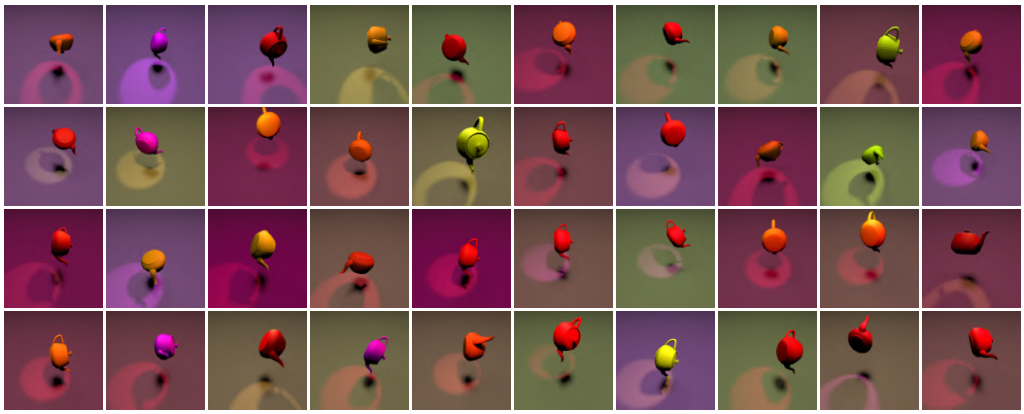


Figure 3: 40 random samples from the marginal distribution of the *Teapot* object class.

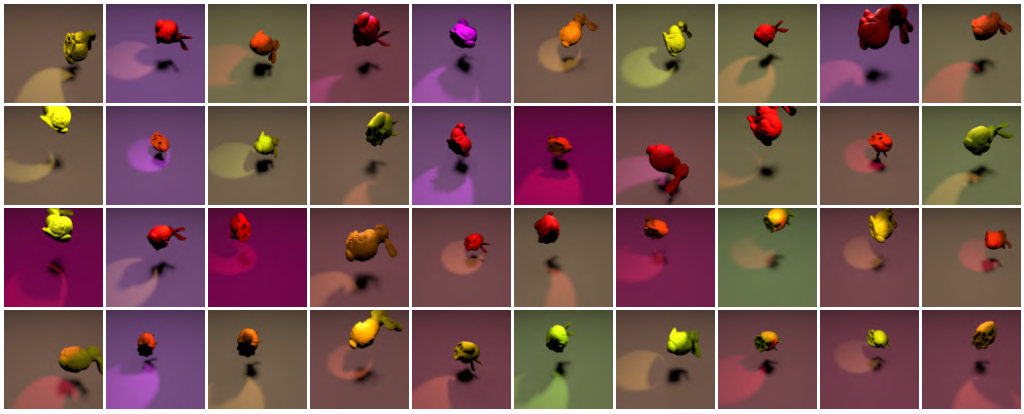


Figure 4: 40 random samples from the marginal distribution of the *Hare* object class.

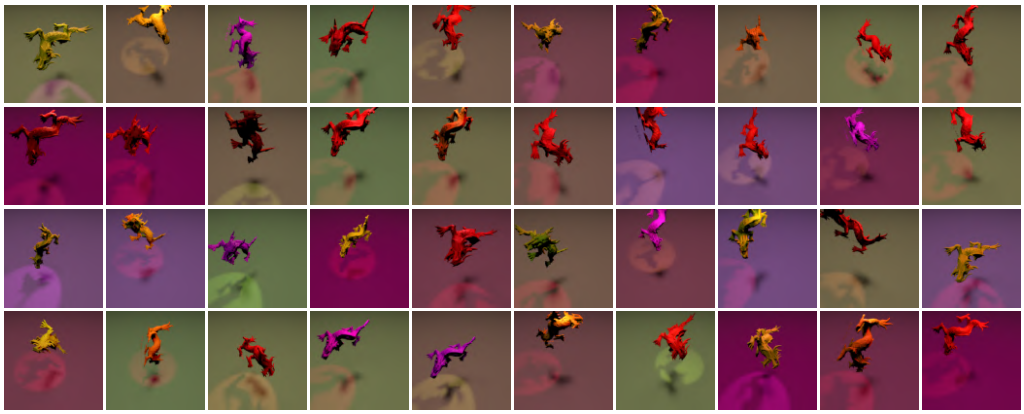


Figure 5: 40 random samples from the marginal distribution of the *Dragon* object class.

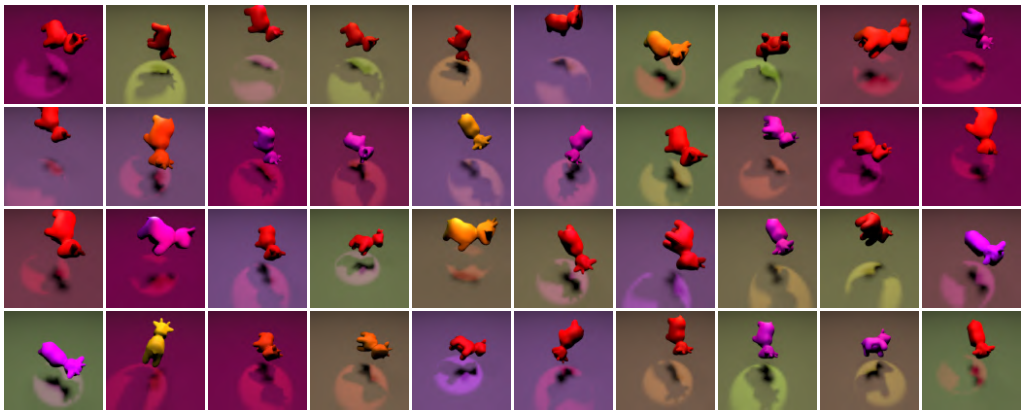


Figure 6: 40 random samples from the marginal distribution of the *Cow* object class.

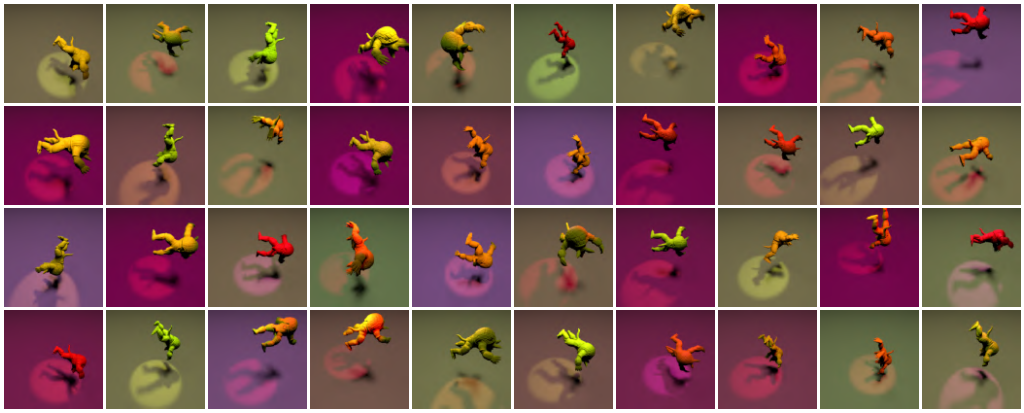


Figure 7: 40 random samples from the marginal distribution of the *Armadillo* object class.

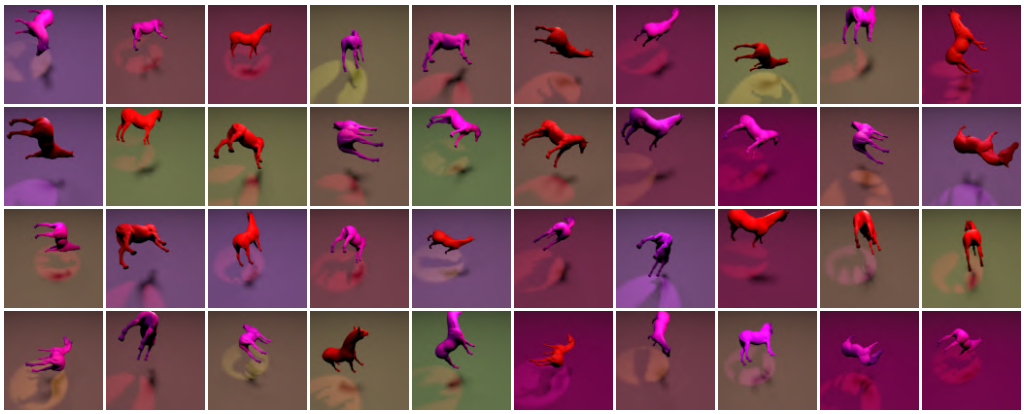


Figure 8: 40 random samples from the marginal distribution of the *Horse* object class.

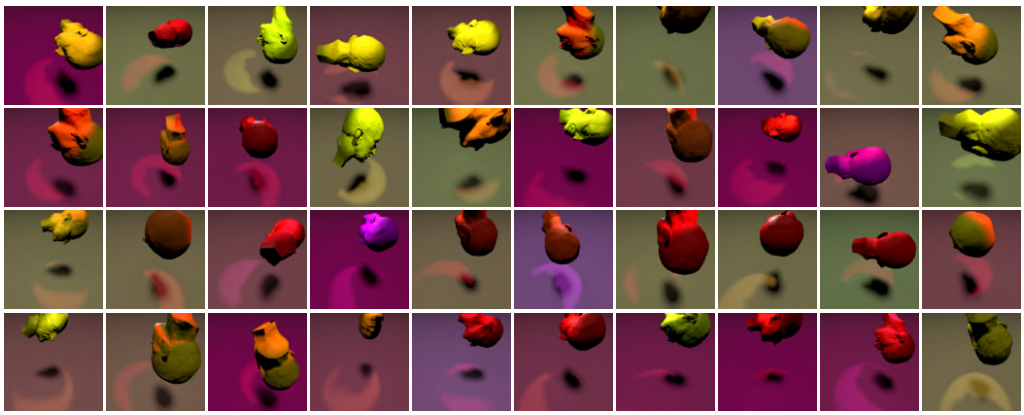


Figure 9: 40 random samples from the marginal distribution of the *Head* object class.

C Additional results

- Appendix C.1 contains numerical experiments, namely linear evaluation & an ablation on $\dim(\hat{\mathbf{c}})$.
- Appendix C.2 contains experiments on *Causal3DIdent*, namely (i) nonlinear & linear evaluation results of the output & intermediate feature representation of `SimCLR` with results for the individual axes of object position & rotation, and (ii) evaluation of `BarlowTwins`.
- Appendix C.3 contains experiments on the *MPI3D-real* dataset [38], namely SimCLR & a supervised sanity check.

C.1 Numerical Data

In Tab. 5, we report mean $\pm$ std. dev. R^2 over 3 random seeds across four generative processes of increasing complexity using *linear* (instead of nonlinear) regression to predict $\mathbf{c}$ from $\hat{\mathbf{c}}$. The block-identification of content can clearly still be seen even if we consider a linear fit.

In Fig. 10, we perform an ablation on $\dim(\hat{\mathbf{c}})$, visualising how varying the dimensionality of the learnt representation affects identifiability of the ground-truth content & style partition. Generally, if $\dim(\hat{\mathbf{c}}) < n_c$, there is insufficient capacity to encode all content, so a lower-dimensional mixture of content is learnt. Conversely, if $\dim(\hat{\mathbf{c}}) > n_c$, the excess capacity is used to encode some style information, as that increases entropy.

Table 5: Results using linear regression for the experiment on numerical data presented in § 5.1

Generative process			$\boldsymbol{R^2}$ **(linear)**	
p(chg.)	**Stat.**	**Cau.**	**Content c**	**Style s**
1.0	✗	✗	**1.00** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✗	✗	**0.99** $\pm$ 0.00	0.00 $\pm$ 0.00
0.75	✓	✗	**0.97** $\pm$ 0.03	0.37 $\pm$ 0.05
0.75	✓	✓	**0.98** $\pm$ 0.01	**0.78** $\pm$ 0.07

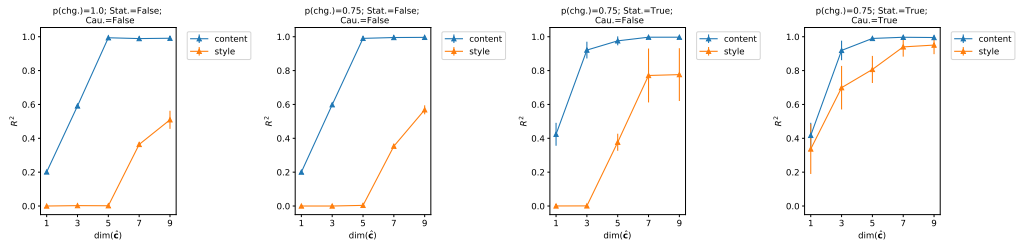


Figure 10: Identifiability of the content & style partition in the numerical experiment as a function of the model latent dimensionality

On Dependence. As can be seen from Tab. 5, the corresponding inset table in § 5.1, and Fig. 10, scores for identifying style increase substantially when statistical dependence within blocks and causal dependence between blocks are included. This finding can be explained as follows.

If we compare the performance for small latent dimensionalities ($\dim(\hat{\mathbf{c}}) < n_c$) between the first two (without) and the third plot (with statistical dependence) of Fig. 10, we observe a significantly higher score in identifying content for the latter (e.g., R^2 of ca. 0.4 vs 0.2 at $\dim(\hat{\mathbf{c}}) = 1$). This suggest that the introduction of statistical dependence between content variables (as well as between style variables, and in how style variables change) in the third plot/row, reduces the effective dimensionality of the ground-truth latents and thus leads to higher content identifiability for the same $\dim(\hat{\mathbf{c}}) < n_c$. Since the R^2 for content is already close to 1 for $\dim(\hat{\mathbf{c}}) = 3$ in the third plot of Fig. 10 (due to the smaller effective dimensionality induced by statistical dependence between $\mathbf{c}$), when $\dim(\hat{\mathbf{c}}) = n_c = 5$ is used (as reported in Tab. 5), excess capacity is used to encode style, leading to a positive R^2.

Regarding causal dependence (i.e., the fourth plot in Fig. 10 and fourth row in Tab. 5), we note that the ground truth dependence between $\mathbf{c}$ and $\mathbf{s}$ is linear, i.e., $p(\mathbf{s}|\mathbf{c})$ is centred at a linear transformation $\mathbf{a} + B\mathbf{c}$ of $\mathbf{c}$, see the data generating process in Appendix D for details. Given that our evaluation

consists of predicting the ground truth $\mathbf{c}$ and $\mathbf{s}$ from the learnt representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$, if we were to block-identify $\mathbf{c}$ according to Defn. 4.1, we should be able to also predict some aspects of $\mathbf{s}$ from $\hat{\mathbf{c}}$, due to the linear dependence between $\mathbf{c}$ and $\mathbf{s}$. This manifests in a relatively large R^2 for $\mathbf{s}$ in the last row of Tab. 5 and the corresponding table in § 5.1.

To summarise, we highlight two main takeaways: (i) when latent dependence is present, this may reduce the effective dimensionality, so that some style is encoded in addition to content unless a smaller representation size is chosen; (ii) even though the learnt representation isolates content in the sense of Defn. 4.1, it may still be predictive of style when content and style are (causally) dependent.

C.2 *Causal3DIdent*

Full version of Tab. 1: In Tab. 6, we a) provide the results for the individual axes of object position & rotation and b) present additional rows omitted from Tab. 1 for space considerations.

Interestingly, we find that the variance across the individual axes is significantly higher for object position than object rotation. If we compare the causal dependence imposed for object position (see Tab. 3) to the causal dependence imposed for object rotation (see Tab. 2), we can observe that the dependence imposed over individual axes is also significantly more variable for position than rotation, i.e., for x the sine nonlinearity is used, for y the cosine nonlinearity is used, while for z, no dependence is imposed.

Regarding the additional rows, we can observe that the composition of image-level rotation & crops yields results quite similar to solely using crops, a relationship which mirrors how transforming the rotation & position latents yields results quite similar to solely transforming the position latents. This suggests that the rotation variables are difficult to disentangle from the position variables in Causal3DIdent, regardless of whether data augmentation or latent transforms are used.

Finally, we can observe that applying image-level rotation in conjunction with small crops & colour distortion does lead to a difference in the encoding, background hue is preserved, while the scores for object position & rotation appear to slightly decrease. When using three augmentations as opposed to two, the effects of the individual augmentations are lessened. While colour distortion discourages the encoding of background hue, both small crops & image-level rotation encourages it, and thus it is preserved when all three augmentations are used. While colour distortion encourages the encoding of object position & rotation, both small crops & image-level rotation discourage it, but as a causal relationship exists between the class variable and said latents, the scores merely decrease, the latents are still for the most part preserved. In reality, where complex interactions between latent variables abound, the effect of data augmentations may be uninterpretable, however with Causal3DIdent, we are able to interpret their effects in the presence of rich visual complexity and causal dependencies, even when applying three distinct augmentations in tandem.

Table 6: Full version of Tab. 1.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.42 ± 0.01	**0.58** ± 0.01	**0.75** ± 0.00	**0.52** ± 0.01	0.17 ± 0.00	0.10 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.36 ± 0.01	0.33 ± 0.01	0.32 ± 0.00
LT: change hues	**1.00** ± 0.00	**0.81** ± 0.02	**0.81** ± 0.02	0.15 ± 0.02	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.02	0.30 ± 0.01	0.30 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.04 ± 0.02	0.03 ± 0.01	0.19 ± 0.02	0.21 ± 0.13	**0.87** ± 0.00	0.09 ± 0.02	**1.00** ± 0.00	0.00 ± 0.00	0.05 ± 0.00	0.02 ± 0.00
DA: crop (small)	0.14 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	0.00 ± 0.00	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LT: change positions	**1.00** ± 0.00	0.01 ± 0.00	0.47 ± 0.01	0.01 ± 0.00	0.00 ± 0.01	0.46 ± 0.02	0.00 ± 0.00	**0.97** ± 0.00	0.30 ± 0.00	0.29 ± 0.00	0.28 ± 0.00
DA: crop (large) + colour distortion	**0.97** ± 0.00	**0.59** ± 0.03	**0.52** ± 0.01	**0.68** ± 0.01	**0.59** ± 0.05	0.28 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	**0.74** ± 0.01	**0.78** ± 0.00	**0.72** ± 0.00
DA: crop (small) + colour distortion	**1.00** ± 0.00	**0.72** ± 0.02	**0.65** ± 0.02	**0.70** ± 0.00	**0.93** ± 0.00	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.03	**0.53** ± 0.00	**0.57** ± 0.01	**0.58** ± 0.01
LT: change positions + hues	**1.00** ± 0.00	0.10 ± 0.10	0.49 ± 0.02	0.06 ± 0.05	0.07 ± 0.08	0.32 ± 0.02	0.00 ± 0.01	0.02 ± 0.03	0.34 ± 0.09	0.34 ± 0.04	0.34 ± 0.08
DA: rotation	0.33 ± 0.06	0.29 ± 0.03	0.11 ± 0.01	0.12 ± 0.04	0.23 ± 0.12	**0.83** ± 0.01	0.30 ± 0.12	**0.99** ± 0.00	0.02 ± 0.01	0.06 ± 0.03	0.07 ± 0.01
LT: change rotations	**1.00** ± 0.00	**0.78** ± 0.01	**0.72** ± 0.03	0.09 ± 0.03	**0.90** ± 0.00	0.41 ± 0.00	0.00 ± 0.00	**0.97** ± 0.00	0.28 ± 0.00	0.28 ± 0.00	0.28 ± 0.00
DA: rotation + colour distortion	**0.59** ± 0.01	**0.63** ± 0.01	**0.57** ± 0.08	**0.54** ± 0.02	0.21 ± 0.01	0.12 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.36 ± 0.03	0.34 ± 0.04	0.30 ± 0.03
LT: change rotations + hues	**1.00** ± 0.00	**0.80** ± 0.02	**0.77** ± 0.01	0.13 ± 0.02	**0.91** ± 0.00	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.00	0.28 ± 0.01	0.28 ± 0.00
DA: rot. + crop (lg)	0.26 ± 0.01	0.03 ± 0.02	0.03 ± 0.01	0.15 ± 0.04	0.04 ± 0.03	**0.84** ± 0.06	0.10 ± 0.01	**1.00** ± 0.00	0.00 ± 0.00	0.04 ± 0.02	0.02 ± 0.00
DA: rot. + crop (sm)	0.15 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	**1.00** ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LT: change rot. + pos.	**1.00** ± 0.00	0.02 ± 0.03	0.48 ± 0.02	0.01 ± 0.01	0.02 ± 0.03	0.49 ± 0.03	0.03 ± 0.02	**0.98** ± 0.00	0.29 ± 0.01	0.28 ± 0.01	0.28 ± 0.01
DA: rot. + crop (lg) + col. dist.	**0.99** ± 0.00	**0.69** ± 0.03	**0.60** ± 0.01	**0.70** ± 0.02	**0.86** ± 0.03	0.28 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	**0.60** ± 0.01	**0.64** ± 0.02	**0.61** ± 0.01
DA: rot. + crop (sm) + col. dist.	**1.00** ± 0.00	**0.61** ± 0.02	**0.59** ± 0.01	**0.64** ± 0.01	**0.82** ± 0.01	0.38 ± 0.00	0.01 ± 0.01	**0.78** ± 0.03	0.44 ± 0.00	0.48 ± 0.02	0.45 ± 0.01
LT: change rot. + pos. + hues	**1.00** ± 0.00	0.20 ± 0.12	**0.50** ± 0.04	0.14 ± 0.11	0.15 ± 0.12	0.32 ± 0.01	0.00 ± 0.00	0.02 ± 0.01	0.33 ± 0.04	0.33 ± 0.02	0.32 ± 0.03

Linear identifiability: In Tab. 7, we present results evaluating all continuous variables with linear regression. While, as expected, R^2 scores are reduced across the board, we can observe that even with a linear fit, the patterns observed in Tab. 6 persist.

Intermediate feature evaluation: In Tab. 8 and Tab. 9, we present evaluation based on the representation from an intermediate layer (i.e., prior to applying a projection layer [20]) with nonlinear and linear regression for the continuous variables, respectively. Note the intermediate layer has an

Table 7: Evaluation results using a linear fit for not only class, but all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.42 ± 0.01	0.37 ± 0.03	0.20 ± 0.16	0.23 ± 0.02	0.01 ± 0.01	0.03 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.13 ± 0.01	0.04 ± 0.01	0.09 ± 0.02
LT: change hues	1.00 ± 0.00	0.72 ± 0.07	0.56 ± 0.04	-0.00 ± 0.00	0.65 ± 0.07	0.29 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.27 ± 0.01	0.26 ± 0.03	0.26 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.00 ± 0.00	0.02 ± 0.00	0.04 ± 0.07	0.08 ± 0.13	0.51 ± 0.05	0.03 ± 0.02	0.20 ± 0.04	0.00 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
DA: crop (small)	0.14 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	0.17 ± 0.05	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00
LT: change positions	1.00 ± 0.00	-0.00 ± 0.00	0.44 ± 0.02	-0.00 ± 0.00	-0.00 ± 0.00	0.29 ± 0.04	0.00 ± 0.00	0.73 ± 0.16	0.26 ± 0.01	0.25 ± 0.03	0.25 ± 0.04
DA: crop (large) + colour distortion	0.97 ± 0.00	0.12 ± 0.02	0.24 ± 0.03	0.21 ± 0.00	0.08 ± 0.03	0.13 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.14 ± 0.04	0.18 ± 0.05	0.22 ± 0.02
DA: crop (small) + colour distortion	1.00 ± 0.00	0.35 ± 0.02	0.50 ± 0.01	0.19 ± 0.03	0.80 ± 0.01	0.28 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	0.29 ± 0.00	0.30 ± 0.00	0.29 ± 0.01
LT: change positions + hues	1.00 ± 0.00	0.00 ± 0.00	0.42 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.27 ± 0.02	-0.00 ± 0.00	-0.00 ± 0.00	0.23 ± 0.07	0.26 ± 0.03	0.25 ± 0.04
DA: rotation	0.33 ± 0.06	0.04 ± 0.04	0.04 ± 0.00	0.02 ± 0.03	0.12 ± 0.08	0.46 ± 0.06	0.06 ± 0.04	0.30 ± 0.13	0.00 ± 0.00	0.04 ± 0.02	0.02 ± 0.00
LT: change rotations	1.00 ± 0.00	0.34 ± 0.21	0.48 ± 0.03	-0.00 ± 0.00	0.60 ± 0.15	0.28 ± 0.00	0.00 ± 0.00	0.59 ± 0.26	0.27 ± 0.01	0.27 ± 0.00	0.27 ± 0.01
DA: rotation + colour distortion	0.59 ± 0.01	0.31 ± 0.02	0.26 ± 0.06	0.25 ± 0.07	0.02 ± 0.00	0.03 ± 0.02	-0.00 ± 0.00	-0.00 ± 0.00	0.07 ± 0.01	0.06 ± 0.01	0.10 ± 0.01
LT: change rotations + hues	1.00 ± 0.00	0.68 ± 0.02	0.57 ± 0.01	-0.00 ± 0.00	0.72 ± 0.10	0.29 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	0.28 ± 0.00	0.28 ± 0.00	0.28 ± 0.00
DA: rot. + crop (lg)	0.26 ± 0.01	-0.00 ± 0.00	0.02 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.59 ± 0.05	0.02 ± 0.01	0.20 ± 0.04	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00
DA: rot. + crop (sm)	0.15 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	0.29 ± 0.21	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00
LT: change rot. + pos.	1.00 ± 0.00	-0.00 ± 0.00	0.45 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.32 ± 0.02	0.00 ± 0.00	0.80 ± 0.09	0.27 ± 0.00	0.27 ± 0.01	0.27 ± 0.01
DA: rot. + crop (lg) + col. dist.	0.99 ± 0.00	0.23 ± 0.04	0.26 ± 0.07	0.26 ± 0.01	0.51 ± 0.14	0.21 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.21 ± 0.04	0.28 ± 0.02	0.22 ± 0.02
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.26 ± 0.02	0.48 ± 0.01	0.21 ± 0.02	0.61 ± 0.01	0.31 ± 0.00	-0.00 ± 0.00	0.34 ± 0.02	0.30 ± 0.00	0.30 ± 0.01	0.29 ± 0.01
LT: change rot. + pos. + hues	1.00 ± 0.00	0.03 ± 0.05	0.46 ± 0.01	0.01 ± 0.01	0.01 ± 0.02	0.29 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	0.27 ± 0.00	0.28 ± 0.01	0.28 ± 0.01

output dimensionality of 100. While it is clear that all R^2 scores are increased across the board, we can notice that certain latents which were discarded in the final layer, were not in an intermediate layer. For example, with "LT: change hues", in the final layer the z-position was discarded ($R^2 = 0.15$ in Tab. 6), inexplicably we may add, as position is content regardless of axis with this latent transformation. But in the intermediate layer, z-position was not discarded ($R^2 = 0.88$ in Tab. 8).

Table 8: Evaluation of an intermediate layer. Logistic regression used for class, kernel ridge regression used for all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.71 ± 0.02	0.68 ± 0.02	0.80 ± 0.01	0.63 ± 0.01	0.25 ± 0.01	0.13 ± 0.00	0.02 ± 0.01	0.01 ± 0.01	0.44 ± 0.01	0.48 ± 0.01	0.39 ± 0.00
LT: change hues	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.88 ± 0.01	0.98 ± 0.00	0.34 ± 0.01	-0.00 ± 0.00	0.20 ± 0.10	0.71 ± 0.02	0.68 ± 0.03	0.68 ± 0.02
DA: crop (large)	0.43 ± 0.03	0.41 ± 0.05	0.35 ± 0.05	0.32 ± 0.04	0.41 ± 0.13	0.88 ± 0.00	0.14 ± 0.03	1.00 ± 0.00	0.03 ± 0.02	0.06 ± 0.01	0.08 ± 0.00
DA: crop (small)	0.20 ± 0.01	0.04 ± 0.05	0.20 ± 0.02	0.01 ± 0.02	0.20 ± 0.03	-0.00 ± 0.00	-0.00 ± 0.00	1.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00
LT: change positions	1.00 ± 0.00	0.78 ± 0.02	0.90 ± 0.01	0.75 ± 0.01	0.59 ± 0.02	0.82 ± 0.01	0.18 ± 0.02	0.99 ± 0.00	0.64 ± 0.02	0.55 ± 0.02	0.56 ± 0.02
DA: crop (large) + colour distortion	1.00 ± 0.00	0.92 ± 0.00	0.83 ± 0.00	0.92 ± 0.00	0.90 ± 0.01	0.29 ± 0.00	0.01 ± 0.01	0.01 ± 0.01	0.87 ± 0.00	0.90 ± 0.00	0.85 ± 0.00
DA: crop (small) + colour distortion	1.00 ± 0.00	0.92 ± 0.00	0.87 ± 0.01	0.90 ± 0.00	0.97 ± 0.00	0.46 ± 0.04	0.02 ± 0.02	0.58 ± 0.12	0.79 ± 0.01	0.83 ± 0.00	0.79 ± 0.00
LT: change positions + hues	1.00 ± 0.00	0.83 ± 0.04	0.90 ± 0.01	0.81 ± 0.04	0.75 ± 0.08	0.42 ± 0.09	0.04 ± 0.02	0.52 ± 0.20	0.72 ± 0.05	0.69 ± 0.07	0.67 ± 0.06
DA: rotation	0.46 ± 0.04	0.35 ± 0.04	0.19 ± 0.02	0.28 ± 0.04	0.34 ± 0.08	0.85 ± 0.01	0.35 ± 0.12	1.00 ± 0.00	0.03 ± 0.01	0.08 ± 0.02	0.10 ± 0.01
LT: change rotations	1.00 ± 0.00	0.97 ± 0.00	0.96 ± 0.01	0.84 ± 0.01	0.98 ± 0.00	0.82 ± 0.01	0.17 ± 0.02	0.99 ± 0.00	0.64 ± 0.02	0.59 ± 0.01	0.60 ± 0.03
DA: rotation + colour distortion	0.87 ± 0.02	0.76 ± 0.01	0.81 ± 0.01	0.71 ± 0.01	0.39 ± 0.08	0.19 ± 0.02	-0.00 ± 0.00	0.02 ± 0.02	0.55 ± 0.03	0.55 ± 0.03	0.48 ± 0.02
LT: change rotations + hues	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.87 ± 0.00	0.99 ± 0.00	0.39 ± 0.05	0.04 ± 0.02	0.37 ± 0.21	0.69 ± 0.01	0.68 ± 0.01	0.68 ± 0.00
DA: rot. + crop (lg)	0.43 ± 0.03	0.38 ± 0.04	0.34 ± 0.02	0.28 ± 0.03	0.30 ± 0.05	0.86 ± 0.04	0.17 ± 0.02	1.00 ± 0.00	0.02 ± 0.00	0.05 ± 0.01	0.10 ± 0.01
DA: rot. + crop (sm)	0.20 ± 0.01	0.07 ± 0.03	0.09 ± 0.10	0.01 ± 0.01	0.20 ± 0.01	-0.00 ± 0.00	-0.00 ± 0.00	1.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00	-0.00 ± 0.00
LT: change rot. + pos.	1.00 ± 0.00	0.81 ± 0.01	0.90 ± 0.01	0.76 ± 0.01	0.67 ± 0.04	0.84 ± 0.01	0.28 ± 0.04	0.99 ± 0.00	0.62 ± 0.02	0.57 ± 0.01	0.55 ± 0.01
DA: rot. + crop (lg) + col. dist.	1.00 ± 0.00	0.92 ± 0.01	0.89 ± 0.00	0.92 ± 0.00	0.95 ± 0.01	0.30 ± 0.00	0.02 ± 0.02	0.18 ± 0.16	0.81 ± 0.00	0.84 ± 0.00	0.79 ± 0.00
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.87 ± 0.00	0.85 ± 0.00	0.87 ± 0.00	0.93 ± 0.00	0.71 ± 0.02	0.33 ± 0.05	0.96 ± 0.00	0.72 ± 0.00	0.75 ± 0.00	0.71 ± 0.00
LT: change rot. + pos. + hues	1.00 ± 0.00	0.84 ± 0.02	0.91 ± 0.01	0.82 ± 0.02	0.78 ± 0.06	0.40 ± 0.01	0.06 ± 0.01	0.50 ± 0.05	0.72 ± 0.04	0.70 ± 0.05	0.67 ± 0.04

Table 9: Evaluation of an intermediate layer. Logistic regression used for class, linear regression used for all continuous variables.

Views generated by	Class	Positions				Hues			Rotations		
		object(x)	object(y)	object(z)	spotlight	object	spotlight	background	object(ϕ)	object(θ)	object(ψ)
DA: colour distortion	0.71 ± 0.02	0.53 ± 0.01	0.70 ± 0.01	0.46 ± 0.01	0.13 ± 0.01	0.11 ± 0.01	-0.01 ± 0.00	0.00 ± 0.00	0.28 ± 0.01	0.19 ± 0.01	0.25 ± 0.01
LT: change hues	1.00 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.60 ± 0.04	0.95 ± 0.00	0.31 ± 0.00	0.01 ± 0.01	0.06 ± 0.04	0.44 ± 0.02	0.41 ± 0.02	0.42 ± 0.00
DA: crop (large)	0.43 ± 0.03	0.18 ± 0.06	0.06 ± 0.01	0.17 ± 0.02	0.19 ± 0.14	0.82 ± 0.02	0.08 ± 0.04	0.98 ± 0.00	0.01 ± 0.00	0.05 ± 0.01	0.05 ± 0.01
DA: crop (small)	0.20 ± 0.01	0.01 ± 0.01	0.03 ± 0.02	0.00 ± 0.01	0.02 ± 0.01	-0.00 ± 0.00	-0.01 ± 0.00	0.99 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.00 ± 0.01
LT: change positions	1.00 ± 0.00	0.49 ± 0.04	0.72 ± 0.03	0.43 ± 0.03	0.19 ± 0.03	0.71 ± 0.02	0.09 ± 0.02	0.98 ± 0.00	0.39 ± 0.01	0.36 ± 0.01	0.35 ± 0.00
DA: crop (large) + colour distortion	1.00 ± 0.00	0.67 ± 0.03	0.56 ± 0.01	0.66 ± 0.02	0.67 ± 0.03	0.28 ± 0.00	-0.01 ± 0.00	0.01 ± 0.01	0.58 ± 0.02	0.61 ± 0.02	0.56 ± 0.01
DA: crop (small) + colour distortion	1.00 ± 0.00	0.76 ± 0.01	0.70 ± 0.02	0.68 ± 0.01	0.90 ± 0.00	0.38 ± 0.03	0.00 ± 0.01	0.39 ± 0.13	0.50 ± 0.02	0.50 ± 0.01	0.49 ± 0.01
LT: change positions + hues	1.00 ± 0.00	0.61 ± 0.09	0.74 ± 0.02	0.51 ± 0.08	0.40 ± 0.15	0.34 ± 0.04	0.02 ± 0.01	0.25 ± 0.22	0.47 ± 0.04	0.40 ± 0.02	0.41 ± 0.03
DA: rotation	0.46 ± 0.04	0.21 ± 0.02	0.10 ± 0.01	0.10 ± 0.02	0.21 ± 0.09	0.77 ± 0.01	0.25 ± 0.11	0.97 ± 0.01	0.02 ± 0.01	0.06 ± 0.02	0.08 ± 0.01
LT: change rotations	1.00 ± 0.00	0.92 ± 0.00	0.88 ± 0.01	0.51 ± 0.02	0.95 ± 0.00	0.70 ± 0.06	0.07 ± 0.02	0.98 ± 0.00	0.36 ± 0.01	0.34 ± 0.00	0.34 ± 0.01
DA: rotation + colour distortion	0.87 ± 0.02	0.60 ± 0.01	0.62 ± 0.03	0.52 ± 0.02	0.23 ± 0.02	0.18 ± 0.02	-0.01 ± 0.00	0.02 ± 0.01	0.33 ± 0.04	0.29 ± 0.01	0.28 ± 0.01
LT: change rotations + hues	1.00 ± 0.00	0.94 ± 0.00	0.92 ± 0.01	0.58 ± 0.01	0.96 ± 0.00	0.33 ± 0.02	0.02 ± 0.01	0.15 ± 0.10	0.40 ± 0.02	0.38 ± 0.01	0.41 ± 0.03
DA: rot. + crop (lg)	0.43 ± 0.03	0.24 ± 0.04	0.08 ± 0.02	0.16 ± 0.03	0.07 ± 0.01	0.80 ± 0.04	0.10 ± 0.01	0.98 ± 0.00	0.01 ± 0.00	0.05 ± 0.01	0.06 ± 0.01
DA: rot. + crop (sm)	0.20 ± 0.01	0.01 ± 0.01	0.03 ± 0.01	-0.00 ± 0.01	0.04 ± 0.01	-0.01 ± 0.00	-0.01 ± 0.00	0.99 ± 0.00	-0.01 ± 0.00	-0.01 ± 0.00	-0.00 ± 0.01
LT: change rot. + pos.	1.00 ± 0.00	0.55 ± 0.05	0.72 ± 0.02	0.44 ± 0.04	0.31 ± 0.08	0.76 ± 0.01	0.14 ± 0.01	0.99 ± 0.00	0.38 ± 0.01	0.35 ± 0.01	0.36 ± 0.02
DA: rot. + crop (lg) + col. dist.	1.00 ± 0.00	0.71 ± 0.01	0.69 ± 0.01	0.69 ± 0.00	0.84 ± 0.03	0.28 ± 0.00	-0.00 ± 0.00	0.07 ± 0.07	0.51 ± 0.01	0.50 ± 0.02	0.51 ± 0.01
DA: rot. + crop (sm) + col. dist.	1.00 ± 0.00	0.66 ± 0.00	0.69 ± 0.01	0.65 ± 0.02	0.83 ± 0.00	0.57 ± 0.03	0.18 ± 0.02	0.89 ± 0.01	0.46 ± 0.01	0.45 ± 0.02	0.44 ± 0.01
LT: change rot. + pos. + hues	1.00 ± 0.00	0.65 ± 0.04	0.75 ± 0.05	0.57 ± 0.03	0.49 ± 0.12	0.35 ± 0.01	0.02 ± 0.01	0.23 ± 0.04	0.48 ± 0.04	0.43 ± 0.01	0.43 ± 0.01

In [20], the value in evaluating an intermediate layer as opposed to a final layer is discussed, where the authors demonstrated that predicting the data augmentations applied during training is significantly more accurate from an intermediate layer as opposed to the final layer, implying that the intermediate layer contains much more information about the transformation applied. Our results suggest a distinct hypothesis, the value in using an intermediate layer as a representation for downstream tasks is not due to preservation of style information, as can be seen, R^2 scores on style variables are not significantly higher in Tab. 8 relative to Tab. 6. The value is in preservation of all content variables, as we can observe certain content variables are discarded in the final layer, but are preserved in an

Table 10: *BarlowTwins* $\lambda = 0.0051$ results: R^2 mean $\pm$ std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	0.48 ± 0.02	**0.51** ± 0.14	0.07 ± 0.01	0.08 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.04
LT: change hues	**1.00** ± 0.00	**0.56** ± 0.20	**0.76** ± 0.07	0.30 ± 0.01	0.00 ± 0.00	0.01 ± 0.00	0.35 ± 0.01
DA: crop (large)	0.17 ± 0.02	0.10 ± 0.03	0.06 ± 0.02	0.29 ± 0.13	0.11 ± 0.05	**0.99** ± 0.00	0.02 ± 0.01
DA: crop (small)	0.15 ± 0.00	0.04 ± 0.02	0.05 ± 0.02	0.02 ± 0.01	0.00 ± 0.01	**1.00** ± 0.00	0.00 ± 0.01
LT: change positions	**0.88** ± 0.00	0.19 ± 0.20	0.05 ± 0.00	**0.50** ± 0.02	0.04 ± 0.01	**0.98** ± 0.00	0.27 ± 0.03
DA: crop (large) + colour distortion	**0.87** ± 0.02	0.49 ± 0.06	0.32 ± 0.03	0.25 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	**0.50** ± 0.02
DA: crop (small) + colour distortion	**0.81** ± 0.01	0.39 ± 0.07	0.42 ± 0.06	0.47 ± 0.04	0.03 ± 0.01	**0.85** ± 0.02	0.30 ± 0.02
LT: change positions + hues	**1.00** ± 0.00	0.28 ± 0.20	0.12 ± 0.05	0.31 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.37 ± 0.06

Table 11: *BarlowTwins* $\lambda = 0.051$ results: R^2 mean $\pm$ std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated.

Views generated by	Class	Positions		Hues			Rotations
		object	spotlight	object	spotlight	background	
DA: colour distortion	**0.52** ± 0.07	0.43 ± 0.18	0.07 ± 0.02	0.10 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.21 ± 0.05
LT: change hues	**1.00** ± 0.00	**0.55** ± 0.24	**0.74** ± 0.02	0.30 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.33 ± 0.02
DA: crop (large)	0.19 ± 0.05	0.08 ± 0.02	0.05 ± 0.01	0.39 ± 0.36	0.08 ± 0.05	**0.96** ± 0.05	0.01 ± 0.02
DA: crop (small)	0.15 ± 0.00	0.05 ± 0.02	0.07 ± 0.02	0.00 ± 0.01	0.01 ± 0.01	**1.00** ± 0.00	0.00 ± 0.00
LT: change positions	**0.89** ± 0.01	0.19 ± 0.20	0.05 ± 0.01	0.48 ± 0.04	0.05 ± 0.02	**0.98** ± 0.00	0.25 ± 0.03
DA: crop (large) + colour distortion	**0.86** ± 0.03	0.40 ± 0.07	0.23 ± 0.02	0.24 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.47 ± 0.04
DA: crop (small) + colour distortion	**0.99** ± 0.01	**0.63** ± 0.03	**0.88** ± 0.01	0.32 ± 0.02	0.00 ± 0.00	0.16 ± 0.13	**0.52** ± 0.03
LT: change positions + hues	**1.00** ± 0.00	0.21 ± 0.22	0.07 ± 0.01	0.30 ± 0.00	0.00 ± 0.00	0.02 ± 0.01	0.46 ± 0.06

intermediate layer. With that being said, our theoretical result applies to the final layer, which is why said results were highlighted in the main paper. The discarding of certain content variables is an empirical phenomenon, likely a consequence of a limited number of negative samples in practice, leading to certain content variables being redundant, or unnecessary, for solving the contrastive objective.

The fact that we can recover certain content variables which appeared discarded in the output from the intermediate layer may suggest that we should be able to decode class. While scores are certainly increased, we do not see such drastic differences in R^2 scores, as was seen above. The drastic difference highlighted above was with regards to latent transformation, for which we always observed class encoded as a content variable. So, unfortunately, using an intermediate layer does not rectify the discrepancy between data augmentations and latent transformations. While latent transformations allow us to better interpret the effect of certain empirical techniques [20], as discussed in the main paper, we cannot make a one-to-one correspondence between data augmentations used in practice and latent transformations.

BarlowTwins: We repeat our analysis from § 5.2 using `BarlowTwins` [128] (instead of `SimCLR`) which, as discussed at the end of § 4.2, is also loosely related to Thm. 4.4. The `BarlowTwins` objective consists of an invariance term, which equates the diagonal elements of the cross-correlation matrix to 1, thereby making the embedding invariant to the distortions applied and a redundancy reduction term, which equates the off-diagonal elements of the cross-correlation matrix to 0, thereby decorrelating the different vector components of the embedding, reducing the redundancy between output units.

In Tab. 10 we train `BarlowTwins` with $\lambda = 0.0051$, the default value for the hyperparameter which weights the redundancy reduction term relative to the invariance term. To confirm the insights are robust to the value of λ,in Tab. 11, we report results with λ increased by an order of magnitude, $\lambda = 0.051$. We find that the results mirror Tab. 1, e.g. colour distortion yields a discarding of hue, crops isolate background hue where the larger the crop, the higher the identifiability of object hue, and crops & colour distortion yield high accuracy in inferring the object class variable.

C.3 *MPI3D-real*

We ran the same experimental setup as in § 5.2 also on the *MPI3D-real* dataset [38] containing > 1 million *real* images with ground-truth annotations of 3D objects being moved by a robotic arm.

Table 12: *MPI3D-real* results: R^2 mean $\pm$ std. dev. over 3 random seeds for dim($\hat{\mathbf{c}}$)$= 5$. DA: data augmentation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$.

Views generated by	object color	object shape	object size	camera height	background color	horizontal axis	vertical axis
DA: colour distortion	0.39 ± 0.01	0.00 ± 0.00	0.16 ± 0.01	**1.00** ± 0.00	0.09 ± 0.15	**0.60** ± 0.06	0.42 ± 0.08
DA: crop (large)	**0.65** ± 0.17	0.01 ± 0.02	0.31 ± 0.03	**1.00** ± 0.00	**1.00** ± 0.00	0.37 ± 0.06	0.08 ± 0.03
DA: crop (small)	0.09 ± 0.02	0.03 ± 0.00	0.19 ± 0.01	**1.00** ± 0.00	**1.00** ± 0.00	0.21 ± 0.02	0.07 ± 0.00
DA: crop (large) + colour distortion	0.34 ± 0.00	0.00 ± 0.00	0.22 ± 0.03	**1.00** ± 0.00	0.39 ± 0.02	**0.54** ± 0.01	0.29 ± 0.01
DA: crop (small) + colour distortion	0.25 ± 0.02	0.00 ± 0.00	0.10 ± 0.01	**1.00** ± 0.00	**0.75** ± 0.16	**0.54** ± 0.01	0.29 ± 0.03

Table 13: **Supervised** *MPI3D-real* results: R^2 mean $\pm$ std. dev. over 3 random seeds. DA: data augmentation. bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$.

Views generated by	object color	object shape	object size	camera height	background color	horizontal axis	vertical axis
Original	**0.90** ± 0.01	0.25 ± 0.02	**0.61** ± 0.02	**0.99** ± 0.00	**0.97** ± 0.01	**1.00** ± 0.00	**1.00** ± 0.00
DA: colour distortion	**0.61** ± 0.01	0.11 ± 0.00	0.47 ± 0.01	**0.98** ± 0.00	**0.93** ± 0.00	**0.99** ± 0.00	**1.00** ± 0.00
DA: crop (large)	**0.82** ± 0.01	0.05 ± 0.01	0.42 ± 0.02	**0.97** ± 0.01	**0.91** ± 0.00	**0.96** ± 0.00	**0.97** ± 0.01
DA: crop (small)	**0.71** ± 0.04	0.01 ± 0.00	0.32 ± 0.02	**0.95** ± 0.00	**0.85** ± 0.01	**0.79** ± 0.02	**0.90** ± 0.01
DA: crop (large) + colour distortion	0.45 ± 0.02	0.02 ± 0.00	0.22 ± 0.00	**0.95** ± 0.01	**0.67** ± 0.01	**0.91** ± 0.00	**0.94** ± 0.00
DA: crop (small) + colour distortion	0.45 ± 0.02	0.00 ± 0.00	0.17 ± 0.02	**0.91** ± 0.02	**0.55** ± 0.03	**0.69** ± 0.01	**0.79** ± 0.08

As *MPI3D-real* contains much lower resolution images (64×64) compared to ImageNet & Causal3DIdent (224×224), we used the standard convolutional encoder from the disentanglement literature [82], and ran a sanity check experiment to verify that by training the same backbone as in our unsupervised experiment with supervised learning, we can recover the ground-truth factors from the augmented views. In Tab. 13, we observe that only five out of seven factors can be consistently inferred, object shape and size are somewhat ambiguous even when observing the original image. Note that while in the self-supervised case, we evaluate by training a nonlinear regression for each ground truth factor separately, in the supervised case, we train a network for all ground truth factors simultaneously from scratch for as many gradient steps as used for learning the self-supervised model.

In Tab. 12, we report the evaluation results in the self-supervised scenario. Subject to the aforementioned caveats, the results show a similar trend as those on *Causal3DIdent*, i.e. with colour distortion, color factors of variation are decoded significantly worse than positional/rotational information.

D Experimental details

Ground-truth generative model. The generative process used in our numerical simulations (§ 5.1) is summarised by the following:

$$
\begin{aligned}
\mathbf{c} \sim p(\mathbf{c}) = \mathcal{N}(0, \Sigma_{\mathbf{c}}), \quad \text{with} \quad \Sigma_{\mathbf{c}} \sim \text{Wishart}_{n_c}(\mathbf{I}, n_c), \\
\mathbf{s}|\mathbf{c} \sim p(\mathbf{s}|\mathbf{c}) = \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_{\mathbf{s}}), \quad \text{with} \quad \Sigma_{\mathbf{s}} \sim \text{Wishart}_{n_s}(\mathbf{I}, n_s), \quad a_i, b_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \\
\tilde{\mathbf{s}}_A|\mathbf{s}_A, A \sim p(\tilde{\mathbf{s}}_A|\mathbf{s}_A) = N(\mathbf{s}_A, \Sigma(A)) \quad \text{with} \quad \Sigma \sim \text{Wishart}_{n_s}(\mathbf{I}, n_s), \\
(\tilde{\mathbf{x}}, \mathbf{x}) = (\mathbf{f}_{\text{MLP}}(\tilde{\mathbf{z}}), \mathbf{f}_{\text{MLP}}(\mathbf{z})),
\end{aligned}
$$

where the set of changing style vectors A is obtained by flipping a (biased) coin with p(chg.) $= 0.75$ for each style dimension independently, and where $\Sigma(A)$ denotes the submatrix of Σ defined by selecting the rows and columns corresponding to subset A.

When we do not allow for *statistical dependence* (Stat.) within blocks of content and style variables, we set the covariance matrices $\Sigma_{\mathbf{c}}$, $\Sigma_{\mathbf{s}}$, and Σ to the identity. When we do not allow for *causal dependence* (Cau.) of style on content, we set $a_i, b_{ij} = 0, \forall i, j$.

For $\mathbf{f}_{\text{MLP}}$, we use a 3-layer MLP with LeakyReLU ($\alpha = 0.2$) activation functions, specified using the same process as used in previous work [54, 55, 129]. For the square weight matrices, we draw $(n_c + n_s) \times (n_c + n_s)$ samples from $U(-1, 1)$, and perform l_2 column normalisation. In addition, to control for invertibility, we re-sample the weight matrices until their condition number is less than or equal to a threshold value. The threshold is pre-computed by sampling $24,975$ weight matrices, and recording the minimum condition number.

Training encoder. Recall that the result of Thm. 4.4 corresponds to minimizing the following functional (5):

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}}\left[\left(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}})\right)^2\right] - H\left(\mathbf{g}(\mathbf{x})\right).$$

Note that InfoNCE [20, 91] (1) can be rewritten as:

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g};\tau,K) = \mathbb{E}_{\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}_{i=1}^K\sim p_{\mathbf{x},\tilde{\mathbf{x}}}}\left[-\sum_{i=1}^K \text{sim}(\mathbf{g}(\mathbf{x})_i,\mathbf{g}(\tilde{\mathbf{x}})_i)/\tau + \log\sum_{j=1}^K \exp\{\text{sim}(\mathbf{g}(\mathbf{x})_i,\mathbf{g}(\tilde{\mathbf{x}})_j)/\tau\}\right]. \quad (32)$$

Thus, if we consider $\tau = 1$, and $\text{sim}(u,v) = -(u-v)^2$,

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g};K) = \mathbb{E}_{\{\mathbf{x}_i,\tilde{\mathbf{x}}_i\}_{i=1}^K\sim p_{\mathbf{x},\tilde{\mathbf{x}}}}\left[\sum_{i=1}^K \left(\mathbf{g}(\mathbf{x})_i - \mathbf{g}(\tilde{\mathbf{x}})_i\right)^2 + \log\sum_{j=1}^K \exp\{-(\mathbf{g}(\mathbf{x})_i - \mathbf{g}(\tilde{\mathbf{x}})_j)^2\}\right] \quad (33)$$

we can approximately match the form of (5). In practice, we use $K = 6,144$.

For $\mathbf{g}$, as in [129], we use a 7-layer MLP with (default) LeakyReLU ($\alpha = 0.01$) activation functions. As the input dimensionality is $(n_c + n_s)$, we consider the following multipliers $[10, 50, 50, 50, 50, 10]$ for the number of hidden units per layer. In correspondence with Thm. 4.4, we set the output dimensionality to n_c.

We train our feature encoder for $300,000$ iterations, using Adam [66] with a learning rate of 10^{-4}.

Causal3DIdent. We here elaborate on details specific to the experiments in § 5.2. We train the feature encoder for $200,000$ iterations using Adam with a learning rate of 10^{-4}. For the encoder we use a ResNet18 [46] architecture followed by a single hidden layer with dimensionality 100 and LeakyReLU activation function using the default (0.01) negative slope. The scores are evaluated on a test set consisting of $25,000$ samples not included in the training set.

Data augmentations. We here specify the parameters for the data augmentations we considered:

- colour distortion: see the paragraph labelled "Color distortion" in Appendix A of [20] for details. We use $s = 1.0$, the default value.
- crop: see the paragraph labelled "Random crop and resize to 224×224" in Appendix A of [20] for details. For small crops, a crop of random size (uniform from 0.08 to 1.0 in area) of the original size is made, which corresponds to what was used in the experiments reported in [20]. For large crops, a crop of random size (uniform from 0.8 to 1.0 in area) of the original size is made.
- rotation: as specified in the captions for Figure 4 & Table 3 in [20], we sample one of $\{0°, 90°, 180°, 270°\}$ uniformly. Note that for the pair, we sample two values without replacement.

A visual overview of the effect of these image-level data augmentations is shown in Fig. 11.

Latent transformations. To generate views via latent transformations (LT) in our experiments on Causal3DIdent (§ 5.2), we proceed as follows.

Let $\mathbf{z}$ refer to the latent corresponding to the original image. For all latents specified to change, we sample $\hat{\mathbf{z}}'$ from a truncated normal distribution constrained to $[-1, 1]$, centered at $\mathbf{z}$, with $\sigma = 1.$. Then, we use nearest-neighbor matching to find the latent $\hat{\mathbf{z}}$ closest to $\hat{\mathbf{z}}'$ (in L^2 distance) for which there exists an image rendering.[15]

Evaluation. Recall that Thm. 4.4 states that $\mathbf{g}$ block-identifies the true content variables in the sense of Defn. 4.1, i.e., there exists an *invertible* function $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

Since this is different from typical evaluation in disentanglement or ICA in that we do not assume independence and do not aim to find a one-to-one correspondence between inferred and ground truth latents, existing metrics, such as MCC [54, 55] or MIG [18], do not apply.

We therefore treat identifying $\mathbf{h}$ as a regression task, which we solve using kernel ridge regression with a Gaussian kernel [88]. Since the Gaussian kernel is universal, this constitutes a nonparametric

[15] see [129] for further details

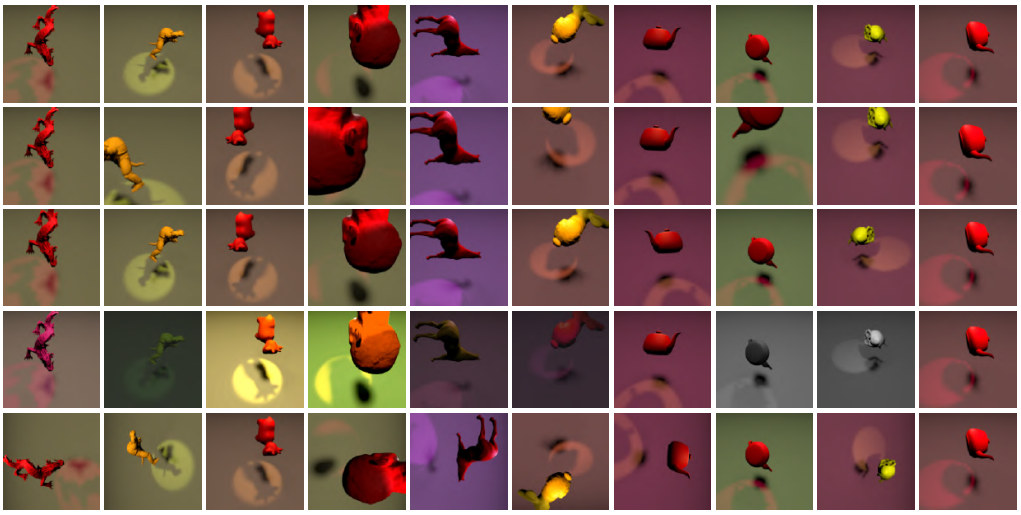


Figure 11: Visual overview of the effect of different data augmentations (DA), applied to 10 representative samples. Rows correspond to *(top to bottom)*: original images, small random crop (+ random flip), large random crop (+ random flip), colour distortion (jitter & drop), and random rotation.

regression technique with universal approximation capabilities, i.e., any nonlinear function can be approximated arbitrarily well given sufficient data.

We sample 4096×10 datapoints from the marginal for evaluation. For kernel ridge regression, we standardize the inputs and targets, and fit the regression model on 4096×5 (distinct) datapoints. We tune the regularization strength α and kernel variance γ by 3-fold cross-validated grid search over the following parameter grids: $\alpha \in [1, 0.1, 0.001, 0.0001]$, $\gamma \in [0.01, 0.22, 4.64, 100]$.

Compute. The experiments in § 5.1 took on the order of 5-10 hours on a single GeForce RTX 2080 Ti GPU. The experiments in § 5.2 on 3DIdent took 28 hours on four GeForce RTX 2080 Ti GPUs. The creation of the Causal3DIdent dataset additionally required approximately 150 hours of compute time on a GeForce RTX 2080 Ti.

2.2 Contrastive Learning Inverts the Data Generating Process

Contrastive Learning Inverts the Data Generating Process

Roland S. Zimmermann [*1,2] **Yash Sharma** [*1,2] **Steffen Schneider** [*1,2,3] **Matthias Bethge** [†1] **Wieland Brendel** [†1]

Abstract

Contrastive learning has recently seen tremendous success in self-supervised learning. So far, however, it is largely unclear why the learned representations generalize so effectively to a large variety of downstream tasks. We here prove that feedforward models trained with objectives belonging to the commonly used InfoNCE family learn to implicitly invert the underlying generative model of the observed data. While the proofs make certain statistical assumptions about the generative model, we observe empirically that our findings hold even if these assumptions are severely violated. Our theory highlights a fundamental connection between contrastive learning, generative modeling, and nonlinear independent component analysis, thereby furthering our understanding of the learned representations as well as providing a theoretical foundation to derive more effective contrastive losses.[1]

1. Introduction

With the availability of large collections of unlabeled data, recent work has led to significant advances in self-supervised learning. In particular, contrastive methods have been tremendously successful in learning representations for visual and sequential data (Logeswaran & Lee, 2018; Wu et al., 2018; Oord et al., 2018; Hénaff, 2020; Tian et al., 2019; Hjelm et al., 2019; Bachman et al., 2019; He et al., 2020a; Chen et al., 2020a; Schneider et al., 2019; Baevski et al., 2020a;b; Ravanelli et al., 2020). While a number of explanations have been provided as to why contrastive learning leads to such informative representations, existing theoretical predictions and empirical observations appear to be at odds with each other (Tian et al., 2019; Bachman et al., 2019; Wu et al., 2020; Saunshi et al., 2019).

In a nutshell, contrastive methods aim to learn representations where related samples are aligned (positive pairs, e.g. augmentations of the same image), while unrelated samples are separated (negative pairs) (Chen et al., 2020a). Intuitively, this leads to invariance to irrelevant details or transformations (by decreasing the distance between positive pairs), while preserving a sufficient amount of information about the input for solving downstream tasks (by increasing the distance between negative pairs) (Tian et al., 2020). This intuition has recently been made more precise by (Wang & Isola, 2020), showing that a commonly used contrastive loss from the InfoNCE family (Gutmann & Hyvärinen, 2012; Oord et al., 2018; Chen et al., 2020a) asymptotically converges to a sum of two losses: an *alignment* loss that pulls together the representations of positive pairs, and a *uniformity* loss that maximizes the entropy of the learned latent distribution.

We show that an encoder learned with a contrastive loss from the InfoNCE family can recover the true generative factors of variation (up to rotations) if the process that generated the data fulfills a few weak statistical assumptions. This theory bridges the gap between contrastive learning, nonlinear independent component analysis (ICA) and generative modeling (see Fig. 1). Our theory reveals implicit assumptions encoded in the InfoNCE objective about the generative process underlying the data. If these assumptions are violated, we show a principled way of deriving alternative contrastive objectives based on assumptions regarding the positive pair distribution. We verify our theoretical findings with controlled experiments, providing evidence that our theory holds true in practice, even if the assumptions on the ground-truth generative model are partially violated.

To the best of our knowledge, our work is the first to analyze under what circumstances representation learning methods used in practice provably represent the data in terms of its underlying factors of variation. Our theoretical and empirical results suggest that the success of contrastive learning in many practical applications is due to an implicit and approximate inversion of the data generating process, which explains why the learned representations are useful in a wide range of downstream tasks.

In summary, our contributions are:

[*]Equal contribution. [†]Joint supervision [1]University of Tübingen, Tübingen, Germany [2]IMPRS for Intelligent Systems, Tübingen, Germany [3]EPFL, Geneva, Switzerland. Correspondence to: Roland S. Zimmermann <roland.zimmermann@uni-tuebingen.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]Online version and code: brendel-group.github.io/cl-ica/

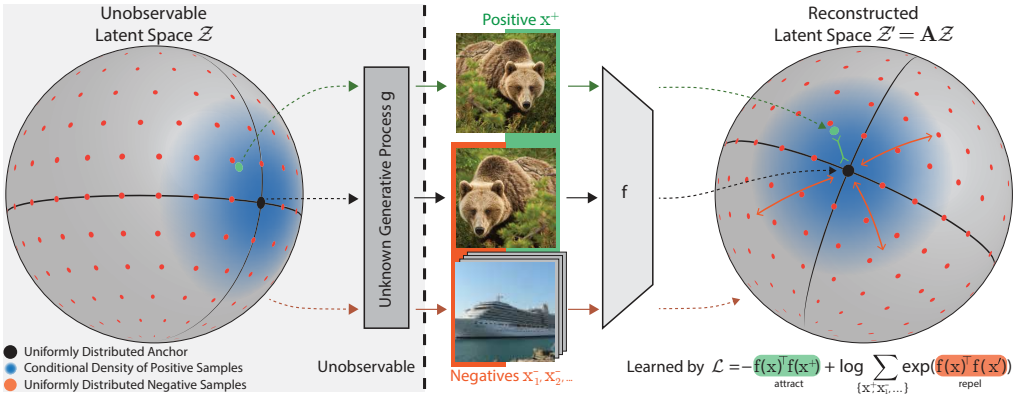


Figure 1. We analyze the setup of contrastive learning, in which a feature encoder f is trained with the InfoNCE objective (Gutmann & Hyvärinen, 2012; Oord et al., 2018; Chen et al., 2020a) using positive samples (green) and negative samples (orange). We assume the observations are generated by an (unknown) injective generative model g that maps unobservable latent variables from a hypersphere to observations in another manifold. Under these assumptions, the feature encoder f implictly learns to invert the ground-truth generative process g up to linear transformations, i.e., $f = \mathbf{A}g^{-1}$ with an orthogonal matrix $\mathbf{A}$, if f minimizes the InfoNCE objective.

- We establish a theoretical connection between the InfoNCE family of objectives, which is commonly used in self-supervised learning, and nonlinear ICA. We show that training with InfoNCE inverts the data generating process if certain statistical assumptions on the data generating process hold.
- We empirically verify our predictions when the assumed theoretical conditions are fulfilled. In addition, we show successful inversion of the data generating process even if these theoretical assumptions are partially violated.
- We build on top of the CLEVR rendering pipeline (Johnson et al., 2017b) to generate a more visually complex disentanglement benchmark, called *3DIdent*, that contains hallmarks of natural environments (shadows, different lighting conditions, a 3D object, etc.). We demonstrate that a contrastive loss derived from our theoretical framework can identify the ground-truth factors of such complex, high-resolution images.

2. Related Work

Contrastive Learning Despite the success of contrastive learning (CL), our understanding of the learned representations remains limited, as existing theoretical explanations yield partially contradictory predictions. One way to theoretically motivate CL is to refer to the InfoMax principle (Linsker, 1988), which corresponds to maximizing the mutual information (MI) between different views (Oord et al., 2018; Bachman et al., 2019; Hjelm et al., 2019; Chen et al., 2020a; Tian et al., 2020). However, as optimizing a tighter bound on the MI can produce worse representations (Tschannen et al., 2020), it is not clear how accurate this motivation describes the behavior of CL.

Another approach aims to explain the success by introducing latent classes (Saunshi et al., 2019). While this theory has some appeal, there exists a gap between empirical observations and its predictions, e.g. the prediction that an excessive number of negative samples decreases performance does not corroborate with empirical results (Wu et al., 2018; Tian et al., 2019; He et al., 2020a; Chen et al., 2020a). However, recent work has suggested some empirical evidence for said theoretical prediction, namely, issues with the commonly used sampling strategy for negative samples, and have proposed ways to mitigate said issues as well (Robinson et al., 2020; Chuang et al., 2020).

More recently, the behavior of CL has been analyzed from the perspective of *alignment* and *uniformity* properties of representations, demonstrating that these two properties are correlated with downstream performance (Wang & Isola, 2020). We build on these results to make a connection to cross-entropy minimization from which we can derive identifiability results.

Nonlinear ICA Independent Components Analysis (ICA) attempts to find the underlying sources for multidimensional data. In the nonlinear case, said sources correspond to a well-defined nonlinear generative model g, which is assumed to be invertible (i.e., injective) (Hyvärinen et al., 2001; Jutten et al., 2010). In other words, nonlinear ICA solves a demixing problem: Given observed data $\mathbf{x} = g(\mathbf{z})$, it aims to find a model f that equals the inverse generative model g^{-1}, which allows for the original sources $\mathbf{z}$ to be recovered.

Hyvärinen et al. (2019) show that the nonlinear demixing problem can be solved as long as the independent compo-

nents are conditionally mutually independent with respect to some auxiliary variable. The authors further provide practical estimation methods for solving the nonlinear ICA problem (Hyvärinen & Morioka, 2016; 2017), similar in spirit to noise contrastive estimation (NCE; Gutmann & Hyvärinen, 2012). Recent work has generalized this contribution to VAEs (Khemakhem et al., 2020a; Locatello et al., 2020; Klindt et al., 2021), as well as (invertible-by-construction) energy-based models (Khemakhem et al., 2020b). We here extend this line of work to more general feed-forward networks trained using InfoNCE (Oord et al., 2018).

In a similar vein, Roeder et al. (2020) build on the work of Hyvärinen et al. (2019) to show that for a model family which includes InfoNCE, distribution matching implies parameter matching. In contrast, we associate the learned latent representation with the ground-truth generative factors, showing under what conditions the data generating process is inverted, and thus, the true latent factors are recovered.

3. Theory

We will show a connection between contrastive learning and identifiability in the form of nonlinear ICA. For this, we introduce a feature encoder f that maps observations $\mathbf{x}$ to representations. We consider the widely used *InfoNCE* loss, which often assumes L^2 normalized representations (Wu et al., 2018; He et al., 2020b; Tian et al., 2019; Bachman et al., 2019; Chen et al., 2020a),

$$\mathcal{L}_{\mathsf{contr}}(f;\tau,M) \quad := \tag{1}$$
$$\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[-\log \frac{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau}}{e^{f(\mathbf{x})^\top f(\tilde{\mathbf{x}})/\tau} + \sum_{i=1}^{M} e^{f(\mathbf{x})^\top f(\mathbf{x}_i^-)/\tau}} \right].$$

Here $M \in \mathbb{Z}_+$ is a fixed number of negative samples, p_{data} is the distribution of all observations and p_{pos} is the distribution of positive pairs. This loss was motivated by the InfoMax principle (Linsker, 1988), and has been shown to be effective by many recent representation learning methods (Logeswaran & Lee, 2018; Wu et al., 2018; Tian et al., 2019; He et al., 2020a; Hjelm et al., 2019; Bachman et al., 2019; Chen et al., 2020a; Baevski et al., 2020b). Our theoretical results also hold for a loss function whose denominator only consists of the second summand across the negative samples (e.g., the SimCLR loss (Chen et al., 2020a)).

In the spirit of existing literature on nonlinear ICA (Hyvärinen & Pajunen, 1999; Harmeling et al., 2003; Sprekeler et al., 2014; Hyvärinen & Morioka, 2016; 2017; Gutmann & Hyvärinen, 2012; Hyvärinen et al., 2019; Khemakhem et al., 2020a), we assume that the observations $\mathbf{x} \in \mathcal{X}$ are generated by an invertible (i.e., injective) generative process $g : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^K$ is the space of observations and $\mathcal{Z} \subseteq \mathbb{R}^N$ with $N \leq K$ denotes the space of latent factors. Influenced by the commonly used feature normalization in InfoNCE, we further assume that $\mathcal{Z}$ is the unit hypersphere $\mathbb{S}^{N-1}$ (see Appx. A.1.1). Additionally, we assume that the ground-truth marginal distribution of the latents of the generative process is uniform and that the conditional distribution (under which positive pairs have high density) is a von Mises-Fisher (vMF) distribution:

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \qquad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \tag{2}$$
$$C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}} = \text{const.}, \quad \mathbf{x} = g(\mathbf{z}), \quad \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}).$$

Given these assumptions, we will show that if f minimizes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$, then f solves the demixing problem, i.e., inverts g up to orthogonal linear transformations.

Our theoretical approach consists of three steps: (1) We demonstrate that $\mathcal{L}_{\mathsf{contr}}$ can be interpreted as the cross-entropy between the (conditional) ground-truth and inferred latent distribution. (2) Next, we show that encoders minimizing $\mathcal{L}_{\mathsf{contr}}$ maintain distance, i.e., two latent vectors with distance α in the ground-truth generative model are mapped to points with the same distance α in the inferred representation. (3) Finally, we leverage distance preservation to show that minimizers of $\mathcal{L}_{\mathsf{contr}}$ invert the generative process up to orthogonal transformations. Detailed proofs are given in Appx. A.1.2.

Additionally, we will present similar results for general convex bodies in $\mathbb{R}^{\mathbb{N}}$ and more general similarity measures, see Sec. 3.3. For this, the detailed proofs are given in Appx. A.2.

3.1. Contrastive learning is related to cross-entropy minimization

From the perspective of nonlinear ICA, we are interested in understanding how the representations $f(\mathbf{x})$ which minimize the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ (defined in Eq. (1)) are related to the ground-truth source signals $\mathbf{z}$. To study this relationship, we focus on the map $h = f \circ g$ between the recovered source signals $h(\mathbf{z})$ and the true source signals $\mathbf{z}$. Note that this is merely for mathematical convenience; it does not necessitate knowledge regarding neither g nor the ground-truth factors during learning (beyond the assumptions stated in the theorems).

A core insight is a connection between the contrastive loss and the cross-entropy between the ground-truth latent distribution and a certain model distribution. For this, we expand the theoretical results obtained by Wang & Isola (2020):

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ converges to the cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive*

loss converges to

$$\lim_{M\to\infty} \mathcal{L}_{\text{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{3}$$

where H is the cross-entropy between the ground-truth conditional distribution p over positive pairs and a conditional distribution q_{h} parameterized by the model f,

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \text{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\, \mathrm{d}\tilde{\mathbf{z}}, \tag{4}$$

where $C_h(\mathbf{z}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appx. A.1.1).

Next, we show that the minimizers h^* of the cross-entropy (4) are isometries in the sense that $\kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}})$ for all $\mathbf{z}$ and $\tilde{\mathbf{z}}$. In other words, they preserve the dot product between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}\exp(\kappa\tilde{\mathbf{z}}^\top\mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau\kappa}$.[2] Consider the conditional distribution q_h parameterized by the model, as defined above in Theorem 1, where the hypothesis class for h (and thus f) is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall\mathbf{z},\tilde{\mathbf{z}} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.*

3.2. Contrastive learning identifies ground-truth factors on the hypersphere

From the strong geometric property of isometry, we can now deduce a key property of the minimizers h^*:

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{Z}' = \mathbb{S}_r^{N-1}$ be the hyperspheres with radius 1 and $r > 0$, respectively. If $h : \mathbb{R}^N \to \mathcal{Z}'$ is differentiable in the vicinity of $\mathcal{Z}$ and its restriction to $\mathcal{Z}$ maintains the dot product up to a constant factor, i.e., $\forall\mathbf{z},\tilde{\mathbf{z}} \in \mathcal{Z} : r^2\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an orthogonal linear transformation scaled by r for all $\mathbf{z} \in \mathcal{Z}$.*

In the last step, we combine the previous propositions to derive our main result: the minimizers of the contrastive loss $\mathcal{L}_{\text{contr}}$ solve the demixing problem of nonlinear ICA up to linear transformations, i.e., they identify the original sources $\mathbf{z}$ for observations $g(\mathbf{z})$ up to orthogonal linear transformations. For a hyperspherical space $\mathcal{Z}$ these correspond to combinations of permutations, rotations and sign flips.

[2]Note that in practice this can be implemented as a learnable rescaling operation as the last operation of the network f.

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). Let the restriction of the mixing function g to $\mathcal{Z}$ be injective and h be differentiable in a vicinity of $\mathcal{Z}$. If the assumed form of q_{h}, as defined above, matches that of p, and if f is differentiable and minimizes the CL loss as defined in Eq. (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is linear, i.e., f recovers the latent sources up to an orthogonal linear transformation and a constant scaling factor.*

Note that we do not assume knowledge of the ground-truth generative model g; we only make assumptions about the conditional and marginal distribution of the latents. On real data, it is unlikely that the assumed model distribution q_{h} can exactly match the ground-truth conditional. We do, however, provide empirical evidence that h is still an affine transformation even if there is a severe mismatch, see Sec. 4.

3.3. Contrastive learning identifies ground-truth factors on convex bodies in $\mathbb{R}^N$

While the previous theoretical results require $\mathcal{Z}$ to be a hypersphere, we will now show a similar theorem for the more general case of $\mathcal{Z}$ being a convex body in $\mathbb{R}^N$. Note that the hyperrectangle $[a_1, b_1] \times \ldots \times [a_N, b_N]$ is an example of such a convex body.

We follow a similar three step proof strategy as for the hyperspherical case before: (1) We begin again by showing that a properly chosen contrastive loss on convex bodies corresponds to the cross-entropy between the ground-truth conditional and a distribution parametrized by the encoder. For this step, we additionally extend the results of Wang & Isola (2020) to this latent space and loss function. (2) Next, we derive that minimizers of the loss function are isometries of the latent space. Importantly, we do not limit ourselves to a specific metric, thus the result is applicable to a family of contrastive objectives. (3) Finally, we show that these minimizers must be affine transformations. For a special family of conditional distributions (rotationally asymmetric generalized normal distributions (Subbotin, 1923)), we can further narrow the class of solutions to permutations and sign-flips. For the detailed proofs, see Appx. A.2.

As earlier, we assume that the ground-truth marginal distribution of the latents is uniform. However, we now assume that the conditional distribution is exponential:

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \qquad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{-\delta(\mathbf{z},\tilde{\mathbf{z}})} \quad \text{with} \quad C_p(\mathbf{z}) := \int e^{-\delta(\mathbf{z},\tilde{\mathbf{z}})}\, \mathrm{d}\tilde{\mathbf{z}}, \quad \mathbf{x} = g(\mathbf{z}), \quad \tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}), \tag{5}$$

where δ is a metric induced by a norm (see Appx. A.2.1).

To reflect the differences between this conditional distribution and the one assumed for the hyperspherical case, we need to introduce an adjusted version of the contrastive loss

in (1):

Definition 1 ($\mathcal{L}_{\delta\text{-contr}}$ objective)**.** Let $\delta : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a metric on $\mathcal{Z}$. We define the general InfoNCE loss, which uses δ as a similarity measure, as

$$\mathcal{L}_{\delta\text{-contr}}(f; \tau, M) := \tag{6}$$
$$\mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\text{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\text{data}}}} \left[-\log \frac{e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau}}{e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} + \sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}_i^-))/\tau}} \right].$$

Note that this is a generalization of the InfoNCE criterion in Eq. (1). In contrast to the objective above, the representations are no longer assumed to be L^2 normalized, and the dot-product is replaced with a more general similarity measure δ.

Analogous to the previously demonstrated case for the hypersphere, for convex bodies $\mathcal{Z}$, minimizers of the adjusted $\mathcal{L}_{\delta\text{-contr}}$ objective solve the demixing problem of nonlinear ICA up to invertible linear transformations:

Theorem 5. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h = f \circ g : \mathcal{Z} \to \mathcal{Z}$, and δ be a metric or a semi-metric (cf. Lemma 1 in Appx. A.2.4), induced by a norm. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as Eq. (5). Let the mixing function g be differentiable and injective. If the assumed form of q_{h} matches that of p, i.e.,*

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau}$$
$$\text{with} \quad C_q(\mathbf{z}) := \int e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{7}$$

and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \to \infty$, we find that $h = f \circ g$ is invertible and affine, i.e., we recover the latent sources up to affine transformations.

Note that the model distribution q_{h}, which is implicitly described by the choice of the objective, must be of the same form as the ground-truth distribution p, i.e., both must be based on the same metric. Thus, identifying different ground-truth conditional distributions requires different contrastive $\mathcal{L}_{\delta\text{-contr}}$ objectives. This result can be seen as a generalized version of Theorem 2, as it is valid for any convex body $\mathcal{Z} \subseteq \mathbb{R}^N$, allowing for a larger variety of conditional distributions.

Finally, under the mild restriction that the ground-truth conditional distribution is based on an L^p similarity measure for $p \geq 1, p \neq 2$, h identifies the ground-truth generative factors up to generalized permutations. A generalized permutation matrix $\mathbf{A}$ is a combination of a permutation and element-wise sign-flips, i.e., $\forall \mathbf{z} : (\mathbf{A}\mathbf{z})_\mathbf{i} = \alpha_\mathbf{i} \mathbf{z}_{\sigma(\mathbf{i})}$ with $\alpha_i = \pm 1$ and σ being a permutation.

Theorem 6. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h : \mathcal{Z} \to \mathcal{Z}$, and δ be an L^α metric or semi-metric (cf. Lemma 1 in Appx. A.2.4) for $\alpha \geq 1, \alpha \neq 2$. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as Eq. (5), and let the mixing function g be differentiable and invertible. If the assumed form of $q_{\text{h}}(\cdot|\mathbf{z})$ matches that of $p(\cdot|\mathbf{z})$, i.e., both use the same metric δ up to a constant scaling factor, and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \to \infty$, we find that $h = f \circ g$ is a composition of input independent permutations, sign flips and rescaling.*

4. Experiments

4.1. Validation of theoretical claim

We validate our theoretical claims under both perfectly matching and violated conditions regarding the ground-truth marginal and conditional distributions. We consider source signals of dimensionality $N = 10$, and sample pairs of source signals in two steps: First, we sample from the marginal $p(\mathbf{z})$. For this, we consider both uniform distributions which match our assumptions and non-uniform distributions (e.g., a normal distribution) which violate them. Second, we generate the positive pair by sampling from a conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z})$. Here, we consider matches with our assumptions on the conditional distribution (von Mises-Fisher for $\mathcal{Z} = \mathbb{S}^{N-1}$) as well as violations (e.g. normal, Laplace or generalized normal distribution for $\mathcal{Z} = \mathbb{S}^{N-1}$). Further, we consider spaces beyond the hypersphere, such as the bounded box (which is a convex body) and the unbounded $\mathbb{R}^N$.

We generate the observations with a multi-layer perceptron (MLP), following previous work (Hyvärinen & Morioka, 2016; 2017). Specifically, we use three hidden layers with leaky ReLU units and random weights; to ensure that the MLP g is invertible, we control the condition number of the weight matrices. For our feature encoder f, we also use an MLP with leaky ReLU units, where the assumed space is denoted by the normalization, or lack thereof, of the encoding. Namely, for the hypersphere (denoted as *Sphere*) and the hyperrectangle (denoted as *Box*) we apply an L^2 and L^∞ normalization, respectively. For flexibility in practice, we parameterize the normalization magnitude of the *Box*, including it as part of the encoder's learnable parameters. On the hypersphere we optimize $\mathcal{L}_{\text{contr}}$ and on the hyperrectangle as well as the unbounded space we optimize $\mathcal{L}_{\delta\text{-contr}}$. For further details, see Appx. A.3.

To test for identifiability up to affine transformations, we fit a linear regression between the ground-truth and recovered sources and report the coefficient of determination (R^2). To test for identifiability up to generalized permutations, we leverage the mean correlation coefficient (MCC), as used

in previous work (Hyvärinen & Morioka, 2016; 2017). For further details, see Appx. A.3.

We evaluate both identifiability metrics for three different model types. First, we ensure that the problem requires nonlinear demixing by considering the identity function for model f, which amounts to scoring the observations against the sources (**Identity Model**). Second, we ensure that the problem is solvable within our model class by training our model f with supervision, minimizing the mean-squared error between $f(g(\mathbf{z}))$ and $\mathbf{z}$ (**Supervised Model**). Third, we fit our model without supervision using a contrastive loss (**Unsupervised Model**).

Tables 1 and 2 show results evaluating identifiability up to affine transformations and generalized permutations, respectively. When assumptions match (see column M.), CL recovers a score close to the empirical upper bound. Mismatches in assumptions on the marginal and conditional do not lead to a significant drop in performance with respect to affine identifiability, but do for permutation identifiability compared to the empirical upper bound. In many practical scenarios, we use the learned representations to solve a downstream task, thus, identifiability up to affine transformations is often sufficient. However, for applications where identification of the individual generative factors is desirable, some knowledge of the underlying generative process is required to choose an appropriate loss function and feature normalization. Interestingly, we find that for convex bodies, we obtain identifiability up to permutation even in the case of a normal conditional, which likely is due to the axis-aligned box geometry of the latent domain. Finally, note that the drop in performance for identifiability up to permutations in the last group of Tab. 2 is a natural consequence of either the ground-truth or the assumed conditional being rotationally symmetric, e.g., a normal distribution, in an unbounded space. Here, rotated versions of the latent space are indistinguishable and, thus, the model cannot align the axes of the reconstruction with that of the ground-truth latent space, resulting in a lower score.

To zoom in on how violations of the uniform marginal assumption influence the identifiability achieved by a model in practice, we perform an ablation on the marginal distribution by interpolating between the theoretically assumed uniform distribution and highly locally concentrated distributions. In particular, we consider two cases: (1) a sphere ($\mathcal{S}^9$) with a vMF marginal around its north pole for different concentration parameters κ; (2) a box ($[0,1]^{10}$) with a normal marginal around the box's center for different standard deviations σ. For both cases, Fig. 2 shows the R^2 score as a function of the concentration κ and $1/\sigma^2$ respectively (black). As a reference, the concentration of the used conditional distribution is highlighted as a dashed line. In addition, we also display the probability mass (0–100%)

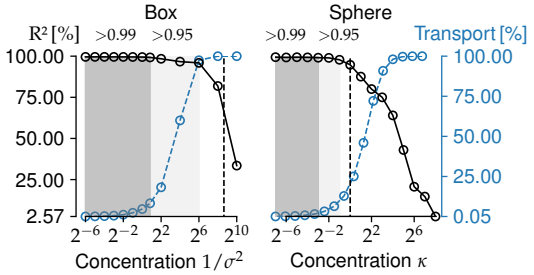


Figure 2. Varying degrees of violation of the uniformity assumption for the marginal distribution. The figure shows the R^2 score measuring identifiability up to linear transformations (black) as well as the difference between the used marginal and assumed uniform distribution in terms of probability mass (blue) as a function of the marginal's concentration. The black dotted line indicates the concentration of the used conditional distribution.

that needs to be moved for converting the used marginal distribution (i.e., vMF or normal) into the assumed uniform marginal distribution (blue) as an intuitive measure of the mismatch (i.e., $\frac{1}{2}\int |p(\mathbf{z}) - p_{\text{uni}}| \, \mathrm{d}\mathbf{z}$). While, we observe significant robustness to mismatch, in both cases, we see performance drop drastically once the marginal distribution is more concentrated than the conditional distribution of positive pairs. In such scenarios, positive pairs are indistinguishable from negative pairs.

4.2. Extensions to image data

Previous studies have demonstrated that representation learning using constrastive learning scales well to complex natural image data (Chen et al., 2020a;b; Hénaff, 2020). Unfortunately, the true generative factors of natural images are inaccessible, thus we cannot evaluate identifiability scores.

We consider two alternatives. First, we evaluate on the recently proposed benchmark *KITTI Masks* (Klindt et al., 2021), which is composed of segmentation masks of natural videos. Second, we contribute a novel benchmark (*3DIdent*; cf. Fig. 3) which features aspects of natural scenes, e.g. a complex 3D object and different lighting conditions, while still providing access to the continuous ground-truth factors. For further details, see Appx. A.4.1. *3DIdent* is available at zenodo.org/record/4502485.

4.2.1. KITTI Masks

KITTI Masks (Klindt et al., 2021) is composed of pedestrian segmentation masks extracted from an autonomous driving vision benchmark KITTI-MOTS (Geiger et al., 2012), with natural shapes and continuous natural transitions. We compare to SlowVAE (Klindt et al., 2021), the state-of-the-art on the considered dataset. In our experiments, we use the same training hyperparameters (for details see Appx. A.3) and (encoder) architecture as Klindt et al. (2021). The positive

Table 1. Identifiability up to affine transformations. Mean ± standard deviation over 5 random seeds. Note that only the first row corresponds to a setting that matches (✓) our theoretical assumptions, while the others show results for violated assumptions (✗; see column *M.*). Note that the identity score only depends on the ground-truth space and the marginal distribution defined for the generative process, while the supervised score additionally depends on the space assumed by the model.

Generative process g			Model f			R^2 Score [%]		
Space	$p(\cdot)$	$p(\cdot\|\cdot)$	Space	$q_h(\cdot\|\cdot)$	M.	Identity	Supervised	Unsupervised
Sphere	Uniform	vMF(κ=1)	Sphere	vMF(κ=1)	✓	66.98 ± 2.79	99.71 ± 0.05	99.42 ± 0.05
Sphere	Uniform	vMF(κ=10)	Sphere	vMF(κ=1)	✗	—"—	—"—	99.86 ± 0.01
Sphere	Uniform	Laplace(λ=0.05)	Sphere	vMF(κ=1)	✗	—"—	—"—	99.91 ± 0.01
Sphere	Uniform	Normal(σ=0.05)	Sphere	vMF(κ=1)	✗	—"—	—"—	99.86 ± 0.00
Box	Uniform	Normal(σ=0.05)	Unbounded	Normal	✗	67.93 ± 7.40	99.78 ± 0.06	99.60 ± 0.02
Box	Uniform	Laplace(λ=0.05)	Unbounded	Normal	✗	—"—	—"—	99.64 ± 0.02
Box	Uniform	Laplace(λ=0.05)	Unbounded	GenNorm(β=3)	✗	—"—	—"—	99.70 ± 0.02
Box	Uniform	Normal(σ=0.05)	Unbounded	GenNorm(β=3)	✗	—"—	—"—	99.69 ± 0.02
Sphere	Normal(σ=1)	Laplace(λ=0.05)	Sphere	vMF(κ=1)	✗	63.37 ± 2.41	99.70 ± 0.07	99.02 ± 0.01
Sphere	Normal(σ=1)	Normal(σ=0.05)	Sphere	vMF(κ=1)	✗	—"—	—"—	99.02 ± 0.02
Unbounded	Laplace(λ=1)	Normal(σ=1)	Unbounded	Normal	✗	62.49 ± 1.65	99.65 ± 0.04	98.13 ± 0.14
Unbounded	Normal(σ=1)	Normal(σ=1)	Unbounded	Normal	✗	63.57 ± 2.30	99.61 ± 0.17	98.76 ± 0.03

Table 2. Identifiability up to generalized permutations, averaged over 5 runs. Note that while Theorem 6 requires the model latent space to be a convex body and $p(\cdot|\cdot) = q_h(\cdot|\cdot)$, we find that empirically either is sufficient. The results are grouped in four blocks corresponding to different types and degrees of violation of assumptions of our theory showing identifiability up to permutations: (1) no violation, violation of the assumptions on either the (2) space or (3) the conditional distribution, or (4) both.

Generative process g			Model f			MCC Score [%]		
Space	$p(\cdot)$	$p(\cdot\|\cdot)$	Space	$q_h(\cdot\|\cdot)$	M.	Identity	Supervised	Unsupervised
Box	Uniform	Laplace(λ=0.05)	Box	Laplace	✓	46.55 ± 1.34	99.93 ± 0.03	98.62 ± 0.05
Box	Uniform	GenNorm(β=3; λ=0.05)	Box	GenNorm(β=3)	✓	—"—	—"—	99.90 ± 0.06
Box	Uniform	Normal(σ=0.05)	Box	Normal	✗	—"—	—"—	99.77 ± 0.01
Box	Uniform	Laplace(λ=0.05)	Box	Normal	✗	—"—	—"—	99.76 ± 0.02
Box	Uniform	GenNorm(β=3; λ=0.05)	Box	Laplace	✗	—"—	—"—	98.80 ± 0.02
Box	Uniform	Laplace(λ=0.05)	Unbounded	Laplace	✗	—"—	99.97 ± 0.03	98.57 ± 0.02
Box	Uniform	GenNorm(β=3; λ=0.05)	Unbounded	GenNorm(β=3)	✗	—"—	—"—	99.85 ± 0.01
Box	Uniform	Normal(σ=0.05)	Unbounded	Normal	✗	—"—	—"—	58.26 ± 3.00
Box	Uniform	Laplace(λ=0.05)	Unbounded	Normal	✗	—"—	—"—	59.67 ± 2.33
Box	Uniform	Normal(σ=0.05)	Unbounded	GenNorm(β=3)	✗	—"—	—"—	43.80 ± 2.15

pairs consist of nearby frames with a time separation $\overline{\Delta t}$.

As argued and shown in Klindt et al. (2021), the transitions in the ground-truth latents between nearby frames is sparse. Unsurprisingly then, Table 3 shows that assuming a Laplace conditional as opposed to a normal conditional in the contrastive loss leads to better identification of the underlying factors of variation. SlowVAE also assumes a Laplace conditional (Klindt et al., 2021) but appears to struggle if the frames of a positive pair are too similar ($\overline{\Delta t} = 0.05s$). This degradation in performance is likely due to the limited expressiveness of the decoder deployed in SlowVAE.

Table 3. **KITTI Masks**. Mean ± standard deviation over 10 random seeds. $\overline{\Delta t}$ indicates the average temporal distance of frames used.

	Model	Model Space	MCC [%]
$\overline{\Delta t} = 0.05s$	SlowVAE	Unbounded	66.1 ± 4.5
	Laplace	Unbounded	77.1 ± 1.0
	Laplace	Box	74.1 ± 4.4
	Normal	Unbounded	58.3 ± 5.4
	Normal	Box	59.9 ± 5.5
$\overline{\Delta t} = 0.15s$	SlowVAE	Unbounded	79.6 ± 5.8
	Laplace	Unbounded	79.4 ± 1.9
	Laplace	Box	80.9 ± 3.8
	Normal	Unbounded	60.2 ± 8.7
	Normal	Box	68.4 ± 6.7

4.2.2. 3DIDENT

Dataset description We build on (Johnson et al., 2017b) and use the Blender rendering engine (Blender Online Com-

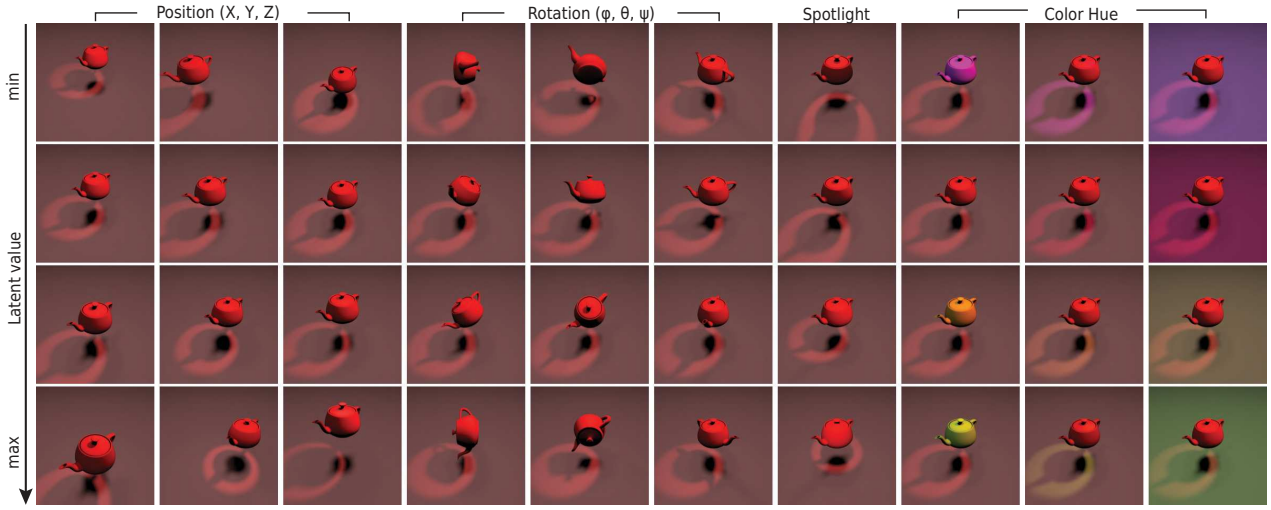


Figure 3. **3DIdent**. Influence of the latent factors $\mathbf{z}$ on the renderings $\mathbf{x}$. Each column corresponds to a traversal in one of the ten latent dimensions while the other dimensions are kept fixed.

munity, 2021) to create visually complex 3D images (see Fig. 3). Each image in the dataset shows a colored 3D object which is located and rotated above a colored ground in a 3D space. Additionally, each scene contains a colored spotlight focused on the object and located on a half-circle around the scene. The observations are encoded with an RGB color space, and the spatial resolution is 224×224 pixels.

The images are rendered based on a 10-dimensional latent, where: (1) three dimensions describe the XYZ position, (2) three dimensions describe the rotation of the object in Euler angles, (3) two dimensions describe the color of the object and the ground of the scene, respectively, and (4) two dimensions describe the position and color of the spotlight. We use the HSV color space to describe the color of the object and the ground with only one latent each by having the latent factor control the hue value. For more details on the dataset see Sec. A.4.

The dataset contains 250 000 observation-latent pairs where the latents are uniformly sampled from the hyperrectangle $\mathcal{Z}$. To sample positive pairs $(\mathbf{z}, \tilde{\mathbf{z}})$ we first sample a value $\tilde{\mathbf{z}}'$ from the data conditional $p(\tilde{\mathbf{z}}'|\mathbf{z})$, and then use nearest-neighbor matching[3] implemented by FAISS (Johnson et al., 2017a) to find the latent $\tilde{\mathbf{z}}$ closest to $\tilde{\mathbf{z}}'$ (in L^2 distance) for which there exists an image rendering. In addition, unlike previous work (Locatello et al., 2019), we create a hold-out test set with 25 000 distinct observation-latent pairs.

Experiments and Results We train a convolutional feature encoder f composed of a ResNet18 architecture (He et al., 2016) and an additional fully-connected layer, with a LeakyReLU nonlinearity as the hidden activation. For more details, see Appx. A.3. Following the same methodology as in Sec. 4.1, i) depending on the assumed space, the output of the feature encoder is normalized accordingly and ii) in addition to the CL models, we also train a supervised model to serve as an upper bound on performance. We consider normal and Laplace distributions for positive pairs. Note, that due to the finite dataset size we only sample from an approximation of these distributions.

As in Tables 1 and 2, the results in Table 4 demonstrate that CL reaches scores close to the topline (supervised) performance, and mismatches between the assumed and ground-truth conditional distribution do not harm the performance significantly. However, if the hypothesis class of the encoder is too restrictive to model the ground-truth conditional distribution, we observe a clear drop in performance, i.e., mapping a box onto a sphere. Note, that this corresponds to the InfoNCE objective for L^2-normalized representations, commonly used for self-supervised representation learning (Wu et al., 2018; He et al., 2020b; Tian et al., 2019; Bachman et al., 2019; Chen et al., 2020a). Finally, the last result shows that leveraging image augmentations (Chen et al., 2020a) as opposed to sampling from a specified conditional distribution of positive pairs $p(\cdot|\cdot)$ results in a performance drop. For details on the experiment, see Appx. Sec. A.3. We explain this with the greater mismatch between the conditional distribution assumed by the model and the conditional distribution induced by the augmentations. In all, we demonstrate validation of our theoretical claims even for generative processes with higher visual complexity than those considered in Sec. 4.1.

[3]We used an Inverted File Index (IVF) with Hierarchical Navigable Small World (HNSW) graph exploration for fast indexing.

Table 4. Identifiability up to affine transformations on the test set of 3DIdent. Mean ± standard deviation over 3 random seeds. As earlier, only the first row corresponds to a setting that matches the theoretical assumptions for linear identifiability; the others show distinct violations. Supervised training with unbounded space achieves scores of $R^2 = (98.67 \pm 0.03)\%$ and MCC $= (99.33 \pm 0.01)\%$. The last row refers to using the image augmentations suggested by Chen et al. (2020a) to generate positive image pairs. For performance on the training set, see Appx. Table 5.

Dataset	Model f			Identity [%]	Unsupervised [%]	
$p(\cdot\|\cdot)$	Space	$q_{\mathrm{h}}(\cdot\|\cdot)$	M.	R^2	R^2	MCC
Normal	Box	Normal	✓	5.25 ± 1.20	96.73 ± 0.10	98.31 ± 0.04
Normal	Unbounded	Normal	✗	—"—	96.43 ± 0.03	54.94 ± 0.02
Laplace	Box	Normal	✗	—"—	96.87 ± 0.08	98.38 ± 0.03
Normal	Sphere	vMF	✗	—"—	65.74 ± 0.01	42.44 ± 3.27
Augm.	Sphere	vMF	✗	—"—	45.51 ± 1.43	46.34 ± 1.59

5. Conclusion

We showed that objectives belonging to the InfoNCE family, the basis for a number of state-of-the-art techniques in self-supervised representation learning, can uncover the true generative factors of variation underlying the observational data. To succeed, these objectives implicitly encode a few weak assumptions about the statistical nature of the underlying generative factors. While these assumptions will likely not be exactly matched in practice, we showed empirically that the underlying factors of variation are identified even if theoretical assumptions are severely violated.

Our theoretical and empirical results suggest that the representations found with contrastive learning implicitly (and approximately) invert the generative process of the data. This could explain why the learned representations are so useful in many downstream tasks. It is known that a decisive aspect of contrastive learning is the right choice of augmentations that form a positive pair. We hope that our framework might prove useful for clarifying the ways in which certain augmentations affect the learned representations, and for finding improved augmentation schemes.

Furthermore, our work opens avenues for constructing more effective contrastive losses. As we demonstrate, imposing a contrastive loss informed by characteristics of the latent space can considerably facilitate inferring the correct semantic descriptors, and thus boost performance in downstream tasks. While our framework already allows for a variety of *conditional* distributions, it is an interesting open question how to adapt it to *marginal* distributions beyond the uniform implicitly encoded in InfoNCE. Also, future work may extend our theoretical framework by incorporating additional assumptions about our visual world, such as compositionality, hierarchy or objectness. Accounting for such inductive biases holds enormous promise in forming the basis for the next generation of self-supervised learning algorithms.

Taken together, we lay a strong theoretical foundation for not only understanding but extending the success of state-of-the-art self-supervised learning techniques.

Author contributions

The project was initiated by WB. RSZ, StS and WB jointly derived the theory. RSZ and YS implemented and executed the experiments. The 3DIdent dataset was created by RSZ with feedback from StS, YS, WB and MB. RSZ, YS, StS and WB contributed to the final version of the manuscript.

Acknowledgements

We thank Muhammad Waleed Gondal, Ivan Ustyuzhaninov, David Klindt, Lukas Schott, Luisa Eck, and Kartik Ahuja for helpful discussions. We thank Bozidar Antic, Shubham Krishna and Jugoslav Stojcheski for ideas regarding the design of 3DIdent. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting RSZ, YS and StS. StS acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (TUE.AI, FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002). WB acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under grant no. BR 6382/1-1 as well as support by Open Philantropy and the Good Ventures Foundation. MB and WB acknowledge funding from the MICrONS program of the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003.

References

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15509–15519, 2019.

Baevski, A., Schneider, S., and Auli, M. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020a.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2021.

Burgess, C. and Kim, H. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Całka, A. Local isometries of compact metric spaces. *Proceedings of the American Mathematical Society*, 85(4): 643–647, 1982.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Chuang, C., Robinson, J., Lin, Y., Torralba, A., and Jegelka, S. Debiased contrastive learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Li, F. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848.

Dittadi, A., Träuble, F., Locatello, F., Wüthrich, M., Agrawal, V., Winther, O., Bauer, S., and Schölkopf, B. On the transfer of disentangled representations in realistic settings. *International Conference on Learning Representations (ICLR)*, 2021.

Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3354–3361. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248074.

Gondal, M. W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 15714–15725, 2019.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

Harmeling, S., Ziehe, A., Kawanabe, M., and Müller, K.-R. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00975.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020b. doi: 10.1109/CVPR42600.2020.00975.

Hénaff, O. J. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4182–4192. PMLR, 2020.

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., and Bengio, Y. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Hyvärinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3765–3773, 2016.

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. Wiley Interscience, 2001.

Hyvärinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017a.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997. IEEE Computer Society, 2017b. doi: 10.1109/CVPR.2017.215.

Jutten, C., Babaie-Zadeh, M., and Karhunen, J. Nonlinear mixtures. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pp. 549–592, 2010.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020a.

Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. *International Conference on Learning Representations (ICLR)*, 2021.

Lamperti, J. et al. On the isometries of certain function-spaces. *Pacific J. Math*, 8(3):459–466, 1958.

Lee, J. M. Smooth manifolds. In *Introduction to Smooth Manifolds*, pp. 606–607. Springer, 2013.

Li, C.-K. and So, W. Isometries of ℓ_p-norm. *The American Mathematical Monthly*, 101(5):452–453, 1994.

Linsker, R. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled

representations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020.

Logeswaran, L. and Lee, H. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Mankiewicz, P. Extension of isometries in normed linear spaces. *Bulletin de l'Academie polonaise des sciences: Serie des sciences mathematiques, astronomiques et physiques*, 20(5):367–+, 1972.

Newell, M. E. *The Utilization of Procedure Models in Digital Image Synthesis.* PhD thesis, The University of Utah, 1975. AAI7529894.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. Multi-task self-supervised learning for robust speech recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6989–6993. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9053569.

Robinson, J., Chuang, C.-Y., Sra, S., and Jegelka, S. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Roeder, G., Metz, L., and Kingma, D. P. On linear identifiability of learned representations. *arXiv preprint arXiv:2007.00810*, 2020.

Ruzhansky, M. and Sugimoto, M. On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences*, 5(1):13–18, 2015.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. A theoretical analysis of contrastive unsupervised representation learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637. PMLR, 2019.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

Sprekeler, H., Zito, T., and Wiskott, L. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

Subbotin, M. F. On the law of frequency of error. *Mat. Sb.*, 31(2):296–301, 1923.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. What makes for good views for contrastive learning, 2020.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Wu, M., Zhuang, C., Yamins, D., and Goodman, N. On the importance of views in unsupervised representation learning. 2020.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3733–3742. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00393.

A. Appendix

A.1. Extended Theory for Hyperspheres

A.1.1. Assumptions

Generative Process Let the generator $g : \mathbb{R}^N \to \mathcal{X}$ with $\mathcal{X} \subseteq \mathbb{R}^K$ and $K \geq N$. Further, let the restriction of g to the space $\mathcal{Z} = \mathbb{S}^{N-1} \subset \mathbb{R}^N$ be injective and g be differentiable in the vicinity of $\mathcal{Z}$. We assume that the marginal distribution $p(\mathbf{z})$ over latent variables $\mathbf{z} \in \mathcal{Z}$ is uniform:

$$p(\mathbf{z}) = \frac{1}{|\mathcal{Z}|}. \tag{8}$$

Further, we assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is a von Mises-Fisher (vMF) distribution

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \tag{9}$$

$$\text{with} \quad C_p := \int e^{\kappa \boldsymbol{\eta}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{10}$$

where κ is a parameter controlling the width of the distribution and $\boldsymbol{\eta}$ is any vector on the hypersphere. Finally, we assume that during training one has access to observations $\mathbf{x}$, which are samples from these distributions transformed by the generator function g.

Model Let $f : \mathcal{X} \to \mathbb{S}_r^{N-1}$, where $\mathbb{S}_r^{N-1}$ denotes a hypersphere with radius r. The parameters of this model are optimized using contrastive learning. We associate a conditional distribution $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through $h = f \circ g$ and

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}$$
$$\text{with} \quad C_q(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \tag{11}$$

where $C_q(\mathbf{z})$ is the partition function and $\tau > 0$ is a scale parameter.

A.1.2. Proofs for Sec. 3

We begin by recalling a result of Wang & Isola (2020), where the authors show an asymptotic relation between the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ and two loss functions, the *alignment* loss $\mathcal{L}_{\mathsf{align}}$ and the *uniformity* loss $\mathcal{L}_{\mathsf{uni}}$:

Proposition A (Asymptotics of $\mathcal{L}_{\mathsf{contr}}$, Wang & Isola, 2020)**.** *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M = \mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uni}}(f;\tau), \tag{12}$$

where

$$\mathcal{L}_{\mathsf{align}}(f;\tau) := -\frac{1}{\tau} \mathop{\mathbb{E}}_{(\tilde{\mathbf{z}},\mathbf{z})\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[(f\circ g)(\mathbf{z})^\top (f\circ g)(\mathbf{z})\right]$$
$$\mathcal{L}_{\mathsf{uni}}(f;\tau) := \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})/\tau}\right]\right]. \tag{13}$$

Proof. See Theorem 1 of Wang & Isola (2020). Note that they originally formulated the losses in terms of observations $\mathbf{x}$ and not in terms of the latent variables $\mathbf{z}$. However, this modified version simplifies notation in the following. □

Based on this result, we show that the contrastive loss $\mathcal{L}_{\mathsf{contr}}$ asymptotically converges to the cross-entropy between the ground-truth conditional p and our assumed model conditional distribution q_h, up to a constant. This is notable, because given the correct model specification for q_h, it is well-known that the cross-entropy is minimized iff $q_h = p$, i.e., the ground-truth conditional distribution and the model distribution will match.

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ converges to the cross-entropy between latent distributions)**.** *If the ground-truth marginal distribution p is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{14}$$

where H is the cross-entropy between the ground-truth conditional distribution p over positive pairs and a conditional distribution q_{h} parameterized by the model f, and $C_h(\mathbf{z}) \in \mathbb{R}^+$ is the partition function of q_{h} (see Appendix A.1.1):

$$q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}$$
$$\textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}. \tag{15}$$

Proof. The cross-entropy between the conditional distributions p and q_h is given by

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{16}$$
$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})} \left[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})\right]\right] \tag{17}$$
$$= \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[-\frac{1}{\tau} h(\tilde{\mathbf{z}})^\top h(\mathbf{z}) + \log C_h(\mathbf{z})\right] \tag{18}$$
$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log C_h(\mathbf{z})\right]. \tag{19}$$

Using the definition of C_h in Eq. (15) we obtain

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{20}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \int_{\mathcal{Z}} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \, \mathrm{d}\tilde{\mathbf{z}}\right]. \tag{21}$$

By assumption the marginal distribution is uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We expand by $|\mathcal{Z}||\mathcal{Z}|^{-1}$ and estimate the integral by sampling from $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, yielding

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{22}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log |\mathcal{Z}| \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\right]\right] \tag{23}$$

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[h(\tilde{\mathbf{z}})^\top h(\mathbf{z})\right] \tag{24}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\right]\right] + \log |\mathcal{Z}|. \tag{25}$$

By inserting the definition $h = f \circ g$,

$$= -\frac{1}{\tau} \mathop{\mathbb{E}}_{\tilde{\mathbf{z}},\mathbf{z}\sim p(\tilde{\mathbf{z}},\mathbf{z})} \left[(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})\right] \tag{26}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{(f\circ g)(\tilde{\mathbf{z}})^\top (f\circ g)(\mathbf{z})/\tau}\right]\right] \tag{27}$$

$$+ \log |\mathcal{Z}|, \tag{28}$$

we can identify the losses introduced in Proposition A,

$$= \mathcal{L}_{\mathsf{align}}(f;\tau) + \mathcal{L}_{\mathsf{uni}}(f;\tau) + \log |\mathcal{Z}|, \tag{29}$$

which recovers the original alignment term and the uniformity term for maximimizing entropy by means of a von Mises-Fisher KDE up to the constant $\log |\mathcal{Z}|$. According to Proposition A this equals

$$= \lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log |Z|, \tag{30}$$

which concludes the proof. □

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} \exp(\kappa \tilde{\mathbf{z}}^\top \mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau\kappa}$.*[4] *Consider the conditional distribution q_h parameterized by the model, as defined above in Theorem 1, where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_h(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.*

[4] Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Proof. By assumption, $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h — in particular, for $h(\mathbf{z}) = \sqrt{\tau\kappa}\mathbf{z}$. The global minimum of the cross-entropy between two distributions is reached if they match by value and have the same support. Thus, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_h(\tilde{\mathbf{z}}|\mathbf{z}). \tag{31}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using that h maps from a unit hypersphere to one with radius $\sqrt{\tau\kappa}$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_h(\mathbf{z}|\mathbf{z}) \tag{32}$$

$$\Leftrightarrow \quad C_p^{-1} e^{\kappa \mathbf{z}^\top \mathbf{z}} = C_h(\mathbf{z})^{-1} e^{h(\mathbf{z})^\top h(\mathbf{z})/\tau} \tag{33}$$

$$\Leftrightarrow \quad C_p^{-1} e^{\kappa} = C_h(\mathbf{z})^{-1} e^{\kappa} \tag{34}$$

$$\Leftrightarrow \quad C_p = C_h. \tag{35}$$

As the normalization constants are identical we get for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} = e^{h(\mathbf{z})^\top h(\tilde{\mathbf{z}})} \Leftrightarrow \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}}). \tag{36}$$

□

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$ and $\mathcal{Z}' = \mathbb{S}_r^{N-1}$ be the hyperspheres with radius 1 and $r > 0$, respectively. If $h : \mathbb{R}^N \to \mathcal{Z}'$ is differentiable in the vicinity of $\mathcal{Z}$ and its restriction to $\mathcal{Z}$ maintains the dot product up to a constant factor, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : r^2 \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an orthogonal linear transformation scaled by r for all $\mathbf{z} \in \mathcal{Z}$.*

Proof. First, we begin with the case $r = 1$. As h maintains the dot product we have:

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}}). \tag{37}$$

We consider the partial derivative w.r.t. $\mathbf{z}$ and obtain:

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \tilde{\mathbf{z}} = \mathbf{J}_h^\top(\mathbf{z}) h(\tilde{\mathbf{z}}). \tag{38}$$

Taking the partial derivative w.r.t. $\tilde{\mathbf{z}}$ yields

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{I} = \mathbf{J}_h^\top(\mathbf{z}) \mathbf{J}_h(\tilde{\mathbf{z}}). \tag{39}$$

We can now conclude

$$\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \mathbf{J}_h(\tilde{\mathbf{z}})^{-1} = \mathbf{J}_h^\top(\mathbf{z}). \tag{40}$$

which implies a constant Jacobian matrix $\mathbf{J}_h(\mathbf{z}) = \mathbf{J}_h$ as the identity holds on all points in $\mathcal{Z}$, and further that the Jacobian $\mathbf{J}_h$ is orthogonal. Hence, $\forall \mathbf{z} \in \mathcal{Z} : h(\mathbf{z}) = \mathbf{J}_h \mathbf{z}$ is an orthogonal linear transformation.

Finally, for $r \neq 1$ we can leverage the previous result by introducing $h'(\mathbf{z}) := h(\mathbf{z})/r$. For h' the previous argument holds, implying that h' is an orthogonal transformation. Therefore, the restriction of h to $\mathcal{Z}$ is an orthogonal linear transformation scaled by r^2. □

Taking all of this together, we can now prove Theorem 2:

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). Let the restriction of the mixing function g to $\mathcal{Z}$ be injective and h be differentiable in a vicinity of $\mathcal{Z}$. If the assumed form of q_h, as defined above, matches that of p, and if f is differentiable and minimizes the CL loss as defined in Eq. (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is linear, i.e., f recovers the latent sources up to an orthogonal linear transformation and a constant scaling factor.*

Proof. As f minimzes the contrastive loss $\mathcal{L}_\mathsf{contr}$ we can apply Theorem 1 to see that f also minimizes the cross-entropy between $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ for any point $\mathbf{z}$ on $\mathcal{Z}$. This means, we can apply Proposition 1 to show that the concatenation $h = f \circ g$ is an isometry with respect to the dot product. Finally, according to Proposition 2, h must then be a composition of an orthogonal linear transformation and a constant scaling factor. Thus, f recovers the latent sources up to orthogonal linear transformations, concluding the proof. □

A.2. Extension of theory to subspaces of $\mathbb{R}^N$

Here, we show how one can generalize the theory above from $\mathcal{Z} = \mathbb{S}^{N-1}$ to $\mathcal{Z} \subseteq \mathbb{R}^N$. Under mild assumptions regarding the ground-truth conditional distribution p and the model distribution q_h, we prove that all minimizers of the cross-entropy between p and q_h are linear functions, if $\mathcal{Z}$ is a convex body. Note that the hyperrectangle $[a_1, b_1] \times \ldots \times [a_N, b_N]$ is an example of such a convex body.

A.2.1. Assumptions

First, we restate the core assumptions for this proof. The main difference to the assumptions for the hyperspherical case above is that we assume different conditional distributions: instead of rotation-invariant von Mises-Fisher distributions, we use translation-invariant distributions (up to restrictions determined by the finite size of the space) of the exponential family.

Generative process Let $g : \mathcal{Z} \to \mathcal{X}$ be an injective function between the two spaces $\mathcal{Z} \subseteq \mathbb{R}^N$ and $\mathcal{X} \subseteq \mathbb{R}^K$ with $K \geq N$ and where $\mathcal{Z}$ is a convex body (e.g., a hyperrectangle). Further, let the marginal distribution be uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$. We assume that the conditional distribution over positive pairs $p(\tilde{\mathbf{z}}|\mathbf{z})$ is an exponential distribution

$$\begin{aligned} p(\tilde{\mathbf{z}}|\mathbf{z}) &= C_p^{-1}(\mathbf{z}) e^{-\lambda \delta(\tilde{\mathbf{z}}, \mathbf{z})} \\ \text{with} \quad C_p(\mathbf{z}) &:= \int e^{-\lambda \delta(\mathbf{z}, \tilde{\mathbf{z}})} \, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{41}$$

where $\lambda > 0$ a parameter controlling the width of the distribution and δ is a (semi-)metric. If δ is a semi-metric, i.e., it does not fulfill the triangle inequality, there must exist a metric δ' such that δ can be written as the composition of a continuously invertible map $j : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $j(0) = 0$ and the metric, i.e., $\delta = j \circ \delta'$. Finally, we assume that during training one has access to samples from both of these distributions.

Note that unlike for the hypersphere, when sampling positive pairs $\mathbf{z}, \tilde{\mathbf{z}} \sim p(\mathbf{z})p(\tilde{\mathbf{z}}|\mathbf{z})$, it is no longer guaranteed that the marginal distributions of $\mathbf{z}$ and $\tilde{\mathbf{z}}$ are the same. When referencing the density functions – or using them in expectation values – $p(\cdot)$ will always denote the same marginal density, no matter if the argument is $\mathbf{z}$ or $\tilde{\mathbf{z}}$. Specifically, $p(\tilde{\mathbf{z}})$ does not refer to $\int p(\mathbf{z})p(\tilde{\mathbf{z}}|\mathbf{z})d\mathbf{z}$.

Model Let $\mathcal{Z}'$ be a subset of $\mathbb{R}^N$ that is a convex body and let $f : \mathcal{X} \to \mathcal{Z}'$ be the model whose parameters are optimized. We associate a conditional distribution $q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ with our model f through

$$\begin{aligned} q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \\ \text{with} \quad C_q(\mathbf{z}) &:= \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{42}$$

where $C_q(\mathbf{z})$ is the partition function and δ is defined above.

A.2.2. Minimizing the cross-entropy

In a first step, we show the analogue of Proposition A for $\mathcal{Z}$ being a convex body:

Proposition 3. *For fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the $\mathcal{L}_{\delta\text{-contr}}$ loss converges to*

$$\begin{aligned} \lim_{M \to \infty} \mathcal{L}_{\delta\text{-contr}}(f; \tau, M) - \log M = \\ \mathcal{L}_{\delta\text{-align}}(f; \tau) + \mathcal{L}_{\delta\text{-uni}}(f; \tau), \end{aligned} \tag{43}$$

where

$$\begin{aligned} \mathcal{L}_{\delta\text{-align}}(f; \tau) &:= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z} \sim p(\mathbf{z}) \\ \tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}}|\mathbf{z})}} \left[\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z})) \right] \\ \mathcal{L}_{\delta\text{-uni}}(f; \tau) &:= \mathop{\mathbb{E}}_{\mathbf{z} \sim p(\mathbf{z})} \left[\log \left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \right] \right) \right], \end{aligned} \tag{44}$$

and $\mathcal{L}_{\delta\text{-contr}}(f; \tau, M)$ is as defined in Eq. (6).

Proof. This proof is adapted from Wang & Isola (2020). By the Continuous Mapping Theorem and the law of large numbers, for any $\mathbf{x}$, $\tilde{\mathbf{x}}$ and $\{ \mathbf{x}_i^- \}_{i=1}^M$ it follows almost surely

$$\begin{aligned}
&\lim_{M\to\infty} \log\left(\frac{1}{M}e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} + \right. \\
&\qquad \left. \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}_i^-))/\tau}\right) \\
&= \log\left(\mathop{\mathbb{E}}_{\mathbf{x}^-\sim p_{\mathsf{data}}}\left[e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right]\right) \\
&= \log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[e^{-\delta(h(\mathbf{z}),h(\tilde{\mathbf{z}}))/\tau}\right]\right),
\end{aligned} \tag{45}$$

where in the last step we expressed the sample $\mathbf{x}$ and negative examples $\mathbf{x}^-$ in terms of their latent factors.

We can now express the limit of the entire loss function as

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f;\tau,M) - \log M \\
&= \frac{1}{\tau}\mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}}}[\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))] \\
&\quad + \lim_{M\to\infty} \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[\log\left(\frac{1}{M}e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} \right.\right. \\
&\qquad \left.\left. + \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right)\right] \\
&= \frac{1}{\tau}\mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}}}[\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))] \\
&\quad + \mathop{\mathbb{E}}_{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}} \left[\lim_{M\to\infty} \log\left(\frac{1}{M}e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))/\tau} \right.\right. \\
&\qquad \left.\left. + \frac{1}{M}\sum_{i=1}^{M} e^{-\delta(f(\mathbf{x}),f(\mathbf{x}_i^-))/\tau}\right)\right].
\end{aligned} \tag{46}$$

Note that as δ is a (semi-)metric, the expression $e^{-\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))}$ is upper-bounded by 1. Hence, according to the Dominated Convergence Theorem one can switch the limit with the expectation value in the second step. Inserting the previous results yields

$$\begin{aligned}
&= \frac{1}{\tau}\mathop{\mathbb{E}}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}}}[\delta(f(\mathbf{x}),f(\tilde{\mathbf{x}}))] \\
&\quad + \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\mathsf{data}}}\left[\log\left(\mathop{\mathbb{E}}_{\mathbf{x}^-\sim p_{\mathsf{data}}}\left[e^{-\delta(f(\mathbf{x}),f(\mathbf{x}^-))/\tau}\right]\right)\right] \\
&= \frac{1}{\tau}\mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\mathbf{z}),h(\tilde{\mathbf{z}}))] \\
&\quad + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})}\left[e^{-\delta(h(\mathbf{z}),h(\tilde{\mathbf{z}}))/\tau}\right]\right)\right] \\
&= \mathcal{L}_{\delta\text{-align}}(f;\tau) + \mathcal{L}_{\delta\text{-uni}}(f;\tau).
\end{aligned} \tag{47}$$

□

Next, we derive a property similar to Theorem 1, which suggests a practical method to find minimizers of the cross-entropy between the ground-truth p and model conditional q_{h}. This property is based on our previously introduced objective function in Eq. (6), which is a modified version of the InfoNCE objective in Eq. (1).

Theorem 3. *Let δ be a semi-metric and $\tau, \lambda > 0$ and let the ground-truth marginal distribution p be uniform. Consider a ground-truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z})\exp(-\lambda\delta(\tilde{\mathbf{z}},\mathbf{z}))$ and the model conditional distribution*

$$\begin{aligned}
q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_h^{-1}(\mathbf{z})e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \\
\textit{with} \quad C_h(\mathbf{z}) &:= \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau}\mathrm{d}\tilde{\mathbf{z}}.
\end{aligned} \tag{48}$$

Then the cross-entropy between p and q_{h} is given by

$$\begin{aligned}
&\lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \\
&\qquad \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}),q_{\mathrm{h}}(\cdot|\mathbf{z})\right],
\end{aligned} \tag{49}$$

which can be implemented by sampling data from the accessible distributions.

Proof. We use the definition of the cross-entropy to write

$$\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}[H(p(\cdot|\mathbf{z}),q_{\mathrm{h}}(\cdot|\mathbf{z})] \tag{50}$$

$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}[\log(q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}))]\right]. \tag{51}$$

We insert the definition of q_{h} and get

$$= -\mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(C_h^{-1}(\mathbf{z})) - \frac{1}{\tau}\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))\right]\right] \tag{52}$$

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}\left[\log(C_h(\mathbf{z})) + \frac{1}{\tau}\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))\right]\right]. \tag{53}$$

As $C_h(\mathbf{z})$ does not depend on $\tilde{\mathbf{z}}$ it can be moved out of the inner expectation value, yielding

$$= \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[\frac{1}{\tau}\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}[\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))] + \log(C_h(\mathbf{z}))\right], \tag{54}$$

which can be written as

$$= \frac{1}{\tau}\mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}}[\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}[\log(C_h(\mathbf{z}))]. \tag{55}$$

Inserting the definition of C_h gives

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} [\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))] \tag{56}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\int e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}} \right) \right]. \tag{57}$$

Next, the second term can be expanded by $1 = |\mathcal{Z}||\mathcal{Z}|^{-1}$, yielding

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} [\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))] \tag{58}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\int \frac{|\mathcal{Z}|}{|\mathcal{Z}|} e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} d\tilde{\mathbf{z}} \right) \right]. \tag{59}$$

Finally, by using that the marginal is uniform, i.e., $p(\mathbf{z}) = |\mathcal{Z}|^{-1}$, this can be simplified as

$$= \frac{1}{\tau} \mathop{\mathbb{E}}_{\substack{\mathbf{z}\sim p(\mathbf{z}) \\ \tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}}|\mathbf{z})}} [\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))] \tag{60}$$

$$+ \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \left[\log \left(\mathop{\mathbb{E}}_{\tilde{\mathbf{z}}\sim p(\tilde{\mathbf{z}})} \left[e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \right] \right) \right] \tag{61}$$

$$+ \log |\mathcal{Z}| \tag{62}$$

$$= \lim_{M\to\infty} \mathcal{L}_{\delta\text{-contr}}(f; \tau, M) - \log M + \log p|\mathcal{Z}|. \tag{63}$$

□

A.2.3. Cross-entropy minimizers are isometries

Now we show a version of Proposition 1, that is generalized from hyperspherical spaces to (subsets of) $\mathbb{R}^N$.

Proposition 4 (Minimizers of the cross-entropy are isometries)**.** *Let δ be a semi-metric. Consider the conditional distributions of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}(\mathbf{z}) \exp(-\delta(\tilde{\mathbf{z}}, \mathbf{z})/\lambda)$ and*

$$\begin{aligned} q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_h^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \\ \text{with} \quad C_h(\mathbf{z}) &:= \int_{\mathcal{Z}} e^{-\delta(h(\tilde{\mathbf{z}}),h(\mathbf{z}))/\tau} \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{64}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ can match for any point $\mathbf{z}$. If h is a minimizer of the cross-entropy $\mathcal{L}_{\mathsf{CE}} = \mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})]$, then h is an isometry, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \lambda\tau\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$.

Proof. Note that $q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z})$ is powerful enough to match $p(\tilde{\mathbf{z}}|\mathbf{z})$ for the correct choice of h, e.g. the identity. The global minimum of cross-entropy between two distributions is reached if they match by value and have the same support. Hence, if p is a regular density, q_{h} will be a regular density, i.e., q_{h} is continuous and has only finite values $0 \leq q_{\mathrm{h}} < \infty$. As the two distributions match, this means

$$p(\tilde{\mathbf{z}}|\mathbf{z}) = q_{\mathrm{h}}(\tilde{\mathbf{z}}|\mathbf{z}). \tag{65}$$

This expression also holds true for $\tilde{\mathbf{z}} = \mathbf{z}$; additionally using the property $\delta(\mathbf{z}, \mathbf{z}) = 0$ yields

$$p(\mathbf{z}|\mathbf{z}) = q_{\mathrm{h}}(\mathbf{z}|\mathbf{z}) \tag{66}$$

$$\Leftrightarrow \quad C_p^{-1}(\mathbf{z}) e^{-\delta(\mathbf{z},\mathbf{z})/\lambda} = C_h^{-1}(\mathbf{z}) e^{-\delta(h(\mathbf{z}),h(\mathbf{z}))/\tau} \tag{67}$$

$$\Leftrightarrow \quad C_p(\mathbf{z}) = C_h(\mathbf{z}). \tag{68}$$

As the normalization constants are identical, we obtain for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$

$$e^{-\delta(\tilde{\mathbf{z}},\mathbf{z})/\lambda} = e^{-\delta(h^*(\tilde{\mathbf{z}}),h^*(\mathbf{z}))/\tau} \tag{69}$$

$$\Leftrightarrow \quad \delta(\tilde{\mathbf{z}}, \mathbf{z}) = \frac{\lambda}{\tau} \delta(h^*(\tilde{\mathbf{z}}), h^*(\mathbf{z})). \tag{70}$$

By introducing a new semi-metric $\delta' := \lambda\tau^{-1}\delta$, we can write this as $\delta(\tilde{\mathbf{z}}, \mathbf{z}) = \delta'(h(\tilde{\mathbf{z}}), h(\mathbf{z}))$, which shows that h is an isometry. If there is no model mismatch, i.e., $\lambda = \tau$, this means $\delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$. □

Note, that this result does not depend on the choice of $\mathcal{Z}$ but just on the class of conditional distributions allowed.

A.2.4. Cross-entropy minimization identifies the ground-truth factors

Before we continue, let us recall a Theorem by Mankiewicz (1972):

Theorem C (Mankiewicz, 1972)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces and let $\mathcal{V}$ be a convex body in $\mathcal{X}$ and $\mathcal{W}$ a convex body in $\mathcal{Y}$. Then every surjective isometry between $\mathcal{V}$ and $\mathcal{W}$ can be uniquely extended to an affine isometry between $\mathcal{X}$ and $\mathcal{Y}$.*

Proof. See Mankiewicz (1972). □

In addition, it is known that isometries on closed spaces are bijective:

Lemma A. *Assume h is an isometry of the closed space $\mathcal{Z}$ into itself, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} : \delta(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))$. Then h is bijective.*

Proof. See Lemma (2.6) in Całka (1982) for surjectivity. We show the injectivity by contradiction. Assume h is not injective. Then we can find a point $\tilde{\mathbf{z}} \neq \mathbf{z}$ where $h(\mathbf{z}) = h(\tilde{\mathbf{z}})$. But then $\delta(\mathbf{z}, \tilde{\mathbf{z}}) > \delta(\mathbf{z}, \mathbf{z})$ and $\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta(h(\mathbf{z}), h(\mathbf{z})) = 0$ by the properties of δ. Hence, h is injective. □

Before continuing, we need to generalize the class of functions we consider as distance measures:

Lemma 1. *Let δ' be a the composition of a continuously invertible function $j : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $j(0) = 0$ and a metric δ, i.e., $\delta' := j \circ \delta$. Then, (i) δ' is a semi-metric and (ii) if a function $h : \mathbb{R}^n \to \mathbb{R}^n$ is an isometry of a space*

with the semi-metric δ', it is also an isometry of the space with the metric δ.

Proof. (i) Let $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$. Per assumption j must be strictly monotonically increasing on $\mathbb{R}_{\geq 0}$. Since δ is a metric it follows $\delta(\mathbf{z}, \tilde{\mathbf{z}}) \geq 0 \Rightarrow \delta'(\mathbf{z}, \tilde{\mathbf{z}}) = j(\delta(\mathbf{z}, \tilde{\mathbf{z}})) \geq 0$, with equality iff $\mathbf{z} = \tilde{\mathbf{z}}$. Furthermore, since δ is a metric it is symmetric in its arguments and, hence, δ' is symmetric in its arguments. Thus, δ' is a semi-metric.

(ii) h is an isometry of a space with the semi-metric δ', allowing to derive that for all $\mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z}$,

$$\delta'(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta'(\mathbf{z}, \tilde{\mathbf{z}}) \tag{71}$$

$$j(\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}}))) = j(\delta(\mathbf{z}, \tilde{\mathbf{z}})) \tag{72}$$

and, applying the inverse j^{-1} which exists by assumption, yields

$$\delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) = \delta(\mathbf{z}, \tilde{\mathbf{z}}), \tag{73}$$

concluding the proof. □

By combining the properties derived before we can show that h is an affine function:

Theorem 4. *Let $\mathcal{Z} = \mathcal{Z}'$ be a convex body in $\mathbb{R}^N$. Let the mixing function g be differentiable and invertible. If the assumed form of q_h as defined in Eq. (42) matches that of p, and if f is differentiable and minimizes the cross-entropy between p and q_h, then we find that $h = f \circ g$ is affine, i.e., we recover the latent sources up to affine transformations.*

Proof. According to Proposition 4 h is an isometry and q_h is a regular probability density function. If the distance δ used in the conditional distributions p and q_h is a semi-metric as in Lemma 1, it follows that h is also an isometry for a proper metric. This also means that h is bijective according to Lemma A. Finally, Theorem C says that h is an affine transformation. □

We use the assumption that the marginal $p(\mathbf{z})$ is uniform, to show

Theorem 5. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h = f \circ g : \mathcal{Z} \to \mathcal{Z}$, and δ be a metric or a semi-metric as defined in Lemma 1. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as (5). Let the mixing function g be differentiable and injective. If the assumed form of q_h matches that of p, i.e.,*

$$\begin{aligned} q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z}) &= C_q^{-1}(\mathbf{z}) e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \\ \text{with} \quad C_q(\mathbf{z}) &:= \int e^{-\delta(h(\tilde{\mathbf{z}}), h(\mathbf{z}))/\tau} \, \mathrm{d}\tilde{\mathbf{z}}, \end{aligned} \tag{74}$$

and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in (6) for $M \to \infty$, we find that $h = f \circ g$ is invertible and affine, i.e., we recover the latent sources up to affine transformations.

Proof. According to Theorem 3 h minimizes the cross-entropy between p and q_h as defined in Eq. (4). Then according to Theorem 4, h is an affine transformation. □

This result can be seen as a generalized version of Theorem 2, as it is valid for any convex body $\mathcal{Z} \subseteq \mathbb{R}^N$ and allows a larger variety of conditional distributions. A missing step is to extend this theory beyond uniform marginal distributions. This will be addressed in future work.

Under some assumptions we can further narrow down possible forms of h, thus, showing that h in fact solves the nonlinear ICA problem only up to permutations and elementwise transformations.

For this, let us first repeat a result from Li & So (1994), that shows an important property of isometric matrices:

Theorem D. *Suppose $1 \leq \alpha \leq \infty$ and $\alpha \neq 2$. An $n \times n$ matrix $\mathbf{A}$ is an isometry of L^α-norm if and only if $\mathbf{A}$ is a generalized permutation matrix, i.e., $\forall \mathbf{z} : (\mathbf{Az})_\mathbf{i} = \alpha_\mathbf{i} \mathbf{z}_{\sigma(\mathbf{i})}$, with $\alpha_i = \pm 2$ and σ being a permutation.*

Proof. See Li & So (1994). Note that this can also be concluded from the Banach-Lamperti Theorem (Lamperti et al., 1958). □

Leveraging this insight, we can finally show:

Theorem 6. *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h : \mathcal{Z} \to \mathcal{Z}$, and δ be an L^α metric for $\alpha \geq 1, \alpha \neq 2$ or the α-th power of such an L^α metric. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as in Eq. (5), and let the mixing function g be differentiable and invertible. If the assumed form of $q_\mathrm{h}(\cdot|\mathbf{z})$ matches that of $p(\cdot|\mathbf{z})$, i.e., both use the same metric δ up to a constant scaling factor, and if f is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \to \infty$ we find that $h = f \circ g$ is a composition of input independent permutations, sign flips and rescalings.*

Proof. First, we prove the case where both conditional distributions use exactly the same metric. By Theorem 5 h is an affine transformation. Moreover, according to Proposition 4 is an isometry. Thus, by Theorem D, h is a generalized permutation matrix, i.e., a composition of permutations and sign flips.

Finally, for the case that δ matches the similarity measure in the ground-truth conditional distribution defined in Eq. (5) (denoted as δ^*) only up to a constant rescaling factor r, we know

$$\begin{aligned} &\forall \mathbf{z}, \tilde{\mathbf{z}} : \delta^*(\mathbf{z}, \tilde{\mathbf{z}}) = \delta(h(\mathbf{z}), h(\tilde{\mathbf{z}})) \\ \Leftrightarrow\ &\delta^*(\mathbf{z}, \tilde{\mathbf{z}}) = \delta^*\left(\frac{1}{r} h(\mathbf{z}), \frac{1}{r} h(\tilde{\mathbf{z}})\right). \end{aligned} \tag{75}$$

Thus, $\frac{1}{r} h$ is a δ^* isometry and the same argument as above holds, concluding the proof. □

Table 5. Identifiability up to affine transformations on the training set of 3DIdent. Mean ± standard deviation over 3 random seeds. As earlier, only the first row corresponds to a setting that matches the theoretical assumptions for linear identifiability; the others show distinct violations. Supervised training with unbounded space achieves scores of $R^2 = (99.98 \pm 0.01)\%$ and MCC $= (99.99 \pm 0.01)\%$. The last row refers to using the SimCLR (Chen et al., 2020a) augmentations to generate positive pairs. The last row refers to using the image augmentations suggested by Chen et al. (2020a) to generate positive image pairs; for details see Sec. A.3. In contrast to Table 4, the scores here are reported on the same data the models were trained on.

Dataset	Model f			Identity [%]	Unsupervised [%]	
$p(\cdot\|\cdot)$	Space	$q_{\mathrm{h}}(\cdot\|\cdot)$	M.	R^2	R^2	MCC
Normal	Box	Normal	✓	5.35 ± 0.72	97.83 ± 0.13	98.85 ± 0.07
Normal	Unbounded	Normal	✗	—"—	97.72 ± 0.02	55.90 ± 2.22
Laplace	Box	Normal	✗	—"—	97.95 ± 0.05	98.94 ± 0.03
Normal	Sphere	vMF	✗	—"—	66.73 ± 0.03	42.72 ± 3.20
Augm.	Sphere	vMF	✗	—"—	45.94 ± 1.80	47.6 ± 1.45

A.3. Experimental details

For the experiments presented in Sec. 4.1 we train our feature encoder for 300 000 iterations with a batch size of 6144 utilizing Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4}. Like Hyvärinen & Morioka (2016; 2017), for the mixing network, we i) use 0.2 for the angle of the negative slope[5], ii) use L^2 normalized weight matrices with minimum condition number of 25 000 uniformly distributed samples. For the encoder, we i) use the default (0.01) negative slope ii) use 6 hidden layers with dimensionality $[N \cdot 10, N \cdot 50, N \cdot 50, N \cdot 50, N \cdot 50, N \cdot 10]$ and iii) initialize the normalization magnitude as 1. We sample 4096 latents from the marginal for evaluation. For MCC (Hyvärinen & Morioka, 2016; 2017) we use the Pearson correlation coefficient[6]; we found there to be no difference with Spearman[7].

For the experiments presented in Sec. 4.2.1, we use the same architecture as the encoder in (Klindt et al., 2021). As in (Klindt et al., 2021), we train for 300 000 iterations with a batch size of 64 utilizing Adam (Kingma & Ba, 2015) with a learning rate of 10^{-4}. For evaluation, as in (Klindt et al., 2021), we use 10 000 samples and the Spearman correlation coefficient.

For the experiments presented in Sec. 4.2.2, we train the feature encoder for 200 000 iterations using Adam with a learning rate of 10^{-4}. For the encoder we use a ResNet18 (He et al., 2016) architecture followed by a single hidden layer with dimensionality $N \cdot 10$ and LeakyReLU activation function using the default (0.01) negative slope. The scores on the training set are evaluated on 10% of the whole training set, 25 000 random samples. The test set consists of 25 000 samples not included in the training set. For the last row of Tab. 4 and Tab. 5 we used the best-working combination of image augmentations found by Chen et al. (2020a) to sample positive pairs. To be precise, we used a random crop and resize operation followed by a color distortion augmentation. The random crops had a uniformly distributed size (between 8% and 100% of the original image area) and a random aspect ration (between 3/4 and 4/3); subsequently, they were resized to the original image dimension (224×224) again. The color distortion operation itself combined color jittering (i.e., random changes of the brightness, contrast, saturation and hue) with color dropping (i.e., random grayscale conversations). We used the same parameters for these augmentations as recommended by Chen et al. (2020a).

The experiments in Sec. 4.1 took on the order of 5-10 hours on a GeForce RTX 2080 Ti GPU, the experiments on KITTI Masks took 1.5 hours on a GeForce RTX 2080 Ti GPU and those on 3DIdent took 28 hours on four GeForce RTX 2080 Ti GPUs. The creation of the 3DIdent dataset additionally required approximately 150 hours of compute time on a GeForce RTX 2080 Ti.

A.4. Details on 3DIdent

We build on the rendering pipeline of Johnson et al. (2017b) and use the Blender engine (Blender Online Community, 2021), as of version 2.91.0, for image rendering. The scenes depicted in the dataset show a rotated and translated object onto which a spotlight is directed. The spotlight is located on a half-circle above the scene and shines down. The scenes can be described by 10 parameters: the position of the object along the X-, Y- and Z-axis, the rotation of the object described by Euler angles (3), the position of the spotlight described by a polar angle, and the hue of the object, the ground and the spotlight. The value range is $[-3, 3]$ for all position parameters, and is $[-\pi/2, \pi/2]$ for the remaining parameters. The parameters are sampled from a 10-dimensional unit hyperrectangle, then rescaled to their corresponding value range. This ensures that the variance

[5] See e.g. https://pytorch.org/docs/stable/generated/torch.nn.LeakyReLU.html

[6] See e.g. https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html

[7] See e.g. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

of the latent factors is the same for all latent dimensions.

To ensure that the generative process is injective, we take two measures: First, we use a non-rotationally symmetric object (Utah tea pot, Newell, 1975), thus the rotation information is unambiguous. Second, we use different levels of color saturation for the object, the spotlight and the ground (1.0, 0.8 and 0.6, respectively), thus the object is always distinguishable from the ground.

A.4.1. Comparison to existing datasets

The proposed dataset contains high-resolution renderings of an object in a 3D scene. It features some aspects of natural scenes, e.g. complex 3D objects, different lighting conditions and continuous variables. Existing benchmarks (Klindt et al., 2021; Burgess & Kim, 2018; Gondal et al., 2019; Dittadi et al., 2021) for disentanglement in 3D scenes differ in important aspects to 3DIdent.

KITTI Masks (Klindt et al., 2021) only enables evaluating identification of the two-dimensional position and scale of the object instance. In addition, the observed segmentation masks are significantly lower resolution than examples in our dataset. 3D Shapes (Burgess & Kim, 2018) and MPI3D (Gondal et al., 2019) are rendered at the same resolution (64×64) as KITTI Masks. Whereas the dataset contributed by (Dittadi et al., 2021) is rendered at $2\times$ that resolution (128×128), our dataset is rendered at $3.5\times$ that resolution (224×224), the resolution at which natural image classification is typically evaluated (Deng et al., 2009). With that being said, we do note that KITTI Masks is unique in containing frames of natural video, and we thus consider it complementary to 3DIdent.

Burgess & Kim (2018), Dittadi et al. (2021), and Gondal et al. (2019) contribute datasets which contain variable object rotations around one, one, and two rotation axes, respectively, while 3DIdent contains variable object rotation around all three rotation axes as well as variable lighting conditions. Furthermore, each of these datasets were generated by sampling latent factors from an equidistant grid, thus only covering a limited number values along each axis of variation, effectively resulting in a highly coarse discretization of naturally continuous variables. As 3DIdent instead samples the latent factors uniformly in the latent space, this better reflects the continuous nature of the latent dimensions.

A.5. Effects of the Uniformity Loss

In previous work, Wang & Isola (2020) showed that a part of the contrastive (InfoNCE) loss — the uniformity loss — effectively ensures that the encoded features are uniformly distributed over a hypersphere. We now show that this part is crucial to ensure that the mapping is bijective. More precisely, we demonstrate that if the distribution of the encoded/reconstructed latents $h(\mathbf{z})$ has the same support as the distribution of $\mathbf{z}$, and both distributions are regular, i.e., their densities are non-zero and finite, then the transformation h is bijective.

First, we focus on the more general case of a map between manifolds:

Proposition 5. *Let $\mathcal{M}, \mathcal{N}$ be simply connected and oriented $\mathcal{C}^1$ manifolds without boundaries and $h : \mathcal{M} \to \mathcal{N}$ be a differentiable map. Further, let the random variable $\mathbf{z} \in \mathcal{M}$ be distributed according to $\mathbf{z} \sim p(\mathbf{z})$ for a regular density function p, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}(\mathbf{z})$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

Proof. We begin by showing by contradiction that the Jacobian determinant of h does not vanish, i.e., $|\det J_h| > 0$:

Suppose that the Jacobian determinant $|\det J_h|$ vanishes for some $\mathbf{z} \in \mathcal{M}$. Then the inverse of the Jacobian determinant goes to infinity at this point and so does the density of $h(\mathbf{z})$ according to the well-known transformation of probability densities. By assumption, both p and $p_{\#h}$ must be regular density functions and, thus, be finite. This contradicts the initial assumption and so the Jacobian determinant $|\det J_h|$ cannot vanish.

Next, we show that the mapping h is proper. Note that a map is called proper if pre-images of compact sets are compact (Ruzhansky & Sugimoto, 2015). Firstly, a continuous mapping between $\mathcal{M}$ and $\mathcal{N}$ is also closed, i.e., pre-images of closed subsets are also closed (Lee, 2013). In addition, it is well-known that continuous functions on compact sets are bounded. Lastly, according to the Heine–Borel theorem, compact subsets of $\mathbb{R}^D$ are closed and bounded. Taken together, this shows that h is proper.

Finally, according to Theorem 2.1 in (Ruzhansky & Sugimoto, 2015) a proper h with non-vanishing Jacobian determinant is bijective, concluding the proof. □

This theorem directly applies to the case of hyperspheres, which are simply connected and oriented manifolds without boundary. This yields:

Corollary 1. *Let $\mathcal{Z}$ be a hypersphere and $h : \mathcal{Z} \to \mathcal{Z}$ be a differentiable map. Further, let the marginal distribution $p(\mathbf{z})$ of the variable $\mathbf{z} \in \mathcal{Z}$ be a regular density function, i.e., $0 < p < \infty$. If the pushforward $p_{\#h}$ of p through h is also a regular density, i.e., $0 < p_{\#h} < \infty$, then h is a bijection.*

Therefore, we can conclude that a loss term ensuring that the encoded features are distributed according to a regular density function, such as the uniformity term, makes the map h bijective and prevents an information loss. Note that this does not assume that the marginal distribution of

the ground-truth latents $p(\mathbf{z})$ is uniform but only that it is regular and non-vanishing.

Note that while the proposition shows that the uniformity loss is sufficient to ensure bijectivity, we can construct counterexamples if its assumptions (like differentiability) are violated even in just a single point. For instance, the requirement of h being fully differentiable is most likely violated in large unregularized neural networks with ReLU nonlinearities. Here, one might need the full contrastive loss to ensure bijectivity of h.

ArXiv Changelog

- Current Version: Thanks to feedback from readers, we fixed a few inconsistencies in our notation. We also added a considerably simplified proof for Proposition 2.
- June 21, 2021: We studied violations of the uniformity assumption in greater details, and added Figure 2. We thank the anonymous reviewers at ICML for their suggestions. This is also the version available in the proceedings of ICML 2021.
- May 25, 2021: Extensions of the theory: We added additional propositions for the effects of the uniformity loss.
- February 17, 2021: First pre-print.

2.3 Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding

TOWARDS NONLINEAR DISENTANGLEMENT IN NATURAL DATA WITH TEMPORAL SPARSE CODING

David Klindt[*]
University of Tübingen
klindt.david@gmail.com

Lukas Schott[*]
University of Tübingen
lukas.schott@bethgelab.org

Yash Sharma[*]
University of Tübingen
yash.sharma@bethgelab.org

Ivan Ustyuzhaninov
University of Tübingen
ivan.ustyuzhaninov@bethgelab.org

Wieland Brendel
University of Tübingen
wieland.brendel@bethgelab.org

Matthias Bethge[‡]
University of Tübingen
matthias.bethge@bethgelab.org

Dylan M Paiton[‡]
University of Tübingen
dylan.paiton@bethgelab.org

ABSTRACT

Disentangling the underlying generative factors from data has so far been limited to carefully constructed scenarios. We propose a path towards natural data by first showing that the statistics of natural data provide enough structure to enable disentanglement, both theoretically and empirically. Specifically, we provide evidence that objects in natural movies undergo transitions that are typically small in magnitude with occasional large jumps, which is characteristic of a temporally sparse distribution. Leveraging this finding we provide a novel proof that relies on a sparse prior on temporally adjacent observations to recover the true latent variables up to permutations and sign flips, providing a stronger result than previous work. We show that equipping practical estimation methods with our prior often surpasses the current state-of-the-art on several established benchmark datasets without any impractical assumptions, such as knowledge of the number of changing generative factors. Furthermore, we contribute two new benchmarks, Natural Sprites and KITTI Masks, which integrate the measured natural dynamics to enable disentanglement evaluation with more realistic datasets. We test our theory on these benchmarks and demonstrate improved performance. We also identify non-obvious challenges for current methods in scaling to more natural domains. Taken together our work addresses key issues in disentanglement research for moving towards more natural settings.

1 INTRODUCTION

Natural scene understanding can be achieved by decomposing the signal into its underlying factors of variation. An intuitive approach for this problem assumes that a visual representation of the world can be constructed via a generative process that receives factors as input and produces natural signals as output (Bengio et al., 2013). This analogy is justified by the fact that our world is composed of distinct entities that can vary independently, but with regularity imposed by physics. What makes the approach appealing is that it formalizes representation learning by directly comparing representations to underlying ground-truth states, as opposed to the indirect evaluation of benchmarking against heuristic downstream tasks (e.g. object recognition). However, the core issue with this approach is *non-identifiability*, which means a set of possible solutions may all appear equally valid to the model, while only one identifies the true generative factors.

Our work is motivated by the question of whether the statistics of natural data will allow for the formulation of an identifiable model. Our core observation that enables us to make progress in

[*‡]Equal contribution. Code: `https://github.com/bethgelab/slow_disentanglement`

addressing this question is that *generative factors of natural data have sparse transitions*. To estimate these generative factors, we compute statistics on measured transitions of area and position for object masks from large-scale, natural, unstructured videos. Specifically, we extracted over 300,000 object segmentation mask transitions from YouTube-VOS (Xu et al., 2018; Yang et al., 2019) and KITTI-MOTS (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016) (discussed in detail in Appendix D). We fit generalized Laplace distributions to the collected data (Eq. 2), which we indicate with orange lines in Fig. 1. We see empirically that all marginal distributions of temporal transitions are highly sparse and that there exist complex dependencies between natural factors (e.g. motion typically affects both position and apparent size). In this study, we focus on the sparse marginals, which we believe constitutes an important advance that sets the stage for solving further issues and eventually applying the technology to real-world problems. With this information at hand, we are able to provide a stronger proof for capturing the underlying generative factors of the data up to permutations and sign flips that is not covered by previous work (Hyvärinen and Morioka, 2016; 2017; Khemakhem et al., 2020a). Thus, we present the first work, to the best of our knowledge, which proposes a theoretically grounded solution that covers the statistics observed in real videos.

Our contributions are: With measurements from unstructured natural video annotations we provide evidence that natural generative factors undergo sparse changes across time. We provide a proof of identifiability that relies on the observed sparse innovations to identify nonlinearly mixed sources up to a permutation and sign-flips, which we then validate with practical estimation methods for empirical comparisons. We leverage the natural scene information to create novel datasets where the latent transitions between frames follow natural statistics. These datasets provide a benchmark to evaluate how well models can uncover the true latent generative factors in the presence of realistic dynamics. We demonstrate improved disentanglement over previous models on existing datasets and our contributed ones with quantitative metrics from both the disentanglement (Locatello et al., 2018) and the nonlinear ICA community (Hyvärinen and Morioka, 2016). We show via numerous visualization techniques that the learned representations for competing models have important differences, even when quantitative metrics suggest that they are performing equally well.

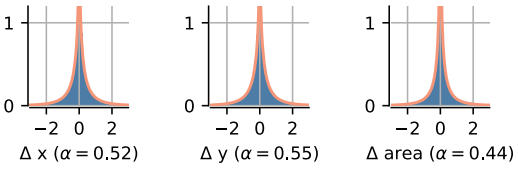


Figure 1: **Statistics of Natural Transitions**. The histograms show distributions over transitions of segmented object masks from natural videos for horizontal and vertical position as well as object size. The red lines indicate fits of generalized Laplace distributions (Eq. 2) with shape value α. Data shown is for object masks extracted from YouTube videos. See Appendix G for 2D marginals and corresponding analysis from the KITTI self-driving car dataset.

2 Related Work – Disentanglement and Nonlinear ICA

Disentangled representation learning has its roots in blind source separation (Cardoso, 1989; Jutten and Herault, 1991) and shares goals with fields such as inverse graphics (Kulkarni et al., 2015; Yildirim et al., 2020; Barron and Malik, 2012) and developing models of invariant neural computation (Hyvärinen and Hoyer, 2000; Wiskott and Sejnowski, 2002; Sohl-Dickstein et al., 2010) (see Bengio et al., 2013, for a review). A disentangled representation would be valuable for a wide variety of machine learning applications, including sample efficiency for downstream tasks (Locatello et al., 2018; Gao et al., 2019), fairness (Locatello et al., 2019; Creager et al., 2019) and interpretability (Bengio et al., 2013; Higgins et al., 2017; Adel et al., 2018). Since there is no agreed upon definition of disentanglement in the literature, we adopt two common measurable criteria: i) each encoding element represents a single generative factor and ii) the values of generative factors are trivially decodable from the encoding (Ridgeway and Mozer, 2018; Eastwood and Williams, 2018).

Uncovering the underlying factors of variation has been a long-standing goal in independent component analysis (ICA) (Comon, 1994; Bell and Sejnowski, 1995), which provides an identifiable solution for disentangling data mixed via an invertible linear generator receiving at most one Gaussian factor as input. Recent unsupervised approaches for nonlinear generators have largely been based on Variational Autoencoders (VAEs) (Kingma and Welling, 2013) and have assumed that the data is independent and identically distributed (*i.i.d.*) (Locatello et al., 2018), even though nonlinear methods that make this *i.i.d.* assumption have been proven to be *non-identifiable* (Hyvärinen and Pajunen,

1999; Locatello et al., 2018). Nonetheless, the bottom-up approach of starting with a nonlinear generator that produces well-controlled data has led to considerable achievements in understanding nonlinear disentanglement in VAEs (Higgins et al., 2017; Burgess et al., 2018; Rolinek et al., 2019; Chen et al., 2018), consolidating ideas from neural computation and machine learning (Khemakhem et al., 2020a), and seeking a principled definition of disentanglement (Ridgeway, 2016; Higgins et al., 2018; Eastwood and Williams, 2018).

Recently, Hyvärinen and colleagues (Hyvärinen and Morioka, 2016; 2017; Hyvärinen et al., 2018) showed that a solution to identifiable nonlinear ICA can be found by assuming that generative factors are conditioned on an additional observed variable, such as past states or the time index itself. This contribution was generalized by Khemakhem et al. (2020a) past the nonlinear ICA domain to any consistent parameter estimation method for deep latent-variable models, including the VAE framework. However, the theoretical assumptions underlying this branch of work do not account for the sparse transitions we observe in the statistics of natural scenes, which we discuss in further detail in appendix F.1.1. Another branch of work requires some form of supervision to demonstrate disentanglement (Szabó et al., 2017; Shu et al., 2019; Locatello et al., 2020). We select two of the above approaches, that are both different in their formulation and state-of-the-art in their respective empirical settings, Hyvärinen and Morioka (2017) and Locatello et al. (2020), for our experiments below. The motivation of our method and dataset contributions is to address the limitations of previous approaches and to enable unsupervised disentanglement learning in more naturalistic scenarios.[1]

The fact that physical processes bind generative factors in temporally adjacent natural video segments has been thoroughly explored for learning in neural networks (Hinton, 1990; Földiák, 1991; Mitchison, 1991; Wiskott and Sejnowski, 2002; Denton and Birodkar, 2017). We propose a method that uses time information in the form of an L_1-*sparse* temporal prior, which is motivated by the natural scene measurements presented above as well as by previous work (Simoncelli and Olshausen, 2001; Olshausen, 2003; Hyvärinen et al., 2003; Cadieu and Olshausen, 2012). Such a prior would intuitively allow for sharp changes in some latent factors, while most other factors remain unchanged between adjacent time-points. Almost all similar methods are variants of slow feature analysis (SFA, Wiskott and Sejnowski, 2002), which measure slowness in terms of the Euclidean (i.e. L_2, or log Gaussian) distance between temporally adjacent encodings. Related to our approach, a probabilistic interpretation of SFA has been previously proposed (Turner and Sahani, 2007), as well as extensions to variational inference (Grathwohl and Wilson, 2016). Additionally, Hashimoto (2003) suggested that a sparse (Cauchy) slowness prior improves correspondence to biological complex cells over the L_2 slowness prior in a two-layer model. However, to the best of our knowledge, an L_1 temporal prior has previously only been used in deep auto-encoder frameworks when applied to semi-supervised tasks (Mobahi et al., 2009; Zou et al., 2012), and was mentioned in Cadieu and Olshausen (2012), who used an L_2 prior, but claimed that an L_1 prior performed similarly on their task. Similar to Hyvärinen et al. (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2018), we only assume that the latent factors are temporally dependent, thus avoiding assuming knowledge of the number of factors where the two observations differ (Shu et al., 2019; Locatello et al., 2020).

Most of the standard datasets for disentanglement (dSprites (Matthey et al., 2017), Cars3D (Reed et al., 2015), SmallNORB (LeCun et al., 2004), Shapes3D (Kim and Mnih, 2018), MPI3D (Gondal et al., 2019)) have been compiled into a disentanglement library (DisLib) by Locatello et al. (2018). However, all of the DisLib datasets are limited in that the data generating process is independent and identically distributed (*i.i.d.*) and all generative factors are assumed to be discrete. In a follow-up study, Locatello et al. (2020) proposed combining pairs of images such that only k factors change, as this matches their modeling assumptions required to prove identifiability. Here, $k \in \mathcal{U}\{1, D-1\}$ and D denotes the number of ground-truth factors, which are then sampled uniformly. We additionally use the measurements from Fig. 1 to construct datasets for evaluating disentanglement that have time transitions which directly correspond to natural dynamics.

[1] As in slow feature analysis, we consider learning from videos without labels as *unsupervised*.

3 THEORY

3.1 GENERATIVE MODEL

We have provided evidence to support the hypothesis that generative factors of natural videos have sparse temporal transitions (see Fig. 1). To model this process, we assume temporally adjacent input pairs $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ coming from a nonlinear generator that maps factors to images $\mathbf{x} = g(\mathbf{z})$, where generative factors are dependent over time:

$$p(\mathbf{z}_t, \mathbf{z}_{t-1}) = p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{z}_{t-1}). \tag{1}$$

Assume the observed data $(\mathbf{x}_t, \mathbf{x}_{t-1})$ comes from the following generative process, where different latent factors are assumed to be independent (cf. Appendix F.2):

$$\mathbf{x} = g(\mathbf{z}), \quad p(\mathbf{z}_{t-1}) = \prod_{i=1}^{d} p(z_{t-1,i}), \quad p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \prod_{i=1}^{d} \frac{\alpha\lambda}{2\Gamma(1/\alpha)} \exp -(\lambda|z_{t,i} - z_{t-1,i}|^{\alpha}), \tag{2}$$

where λ is the distribution rate, $p(\mathbf{z}_{t-1})$ is a factorized Gaussian prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (as in Kingma and Welling, 2013) and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ is a factorized generalized Laplace distribution (Subbotin, 1923) with shape parameter α, which determines the shape and especially the kurtosis of the function.[2] Intuitively, smaller α implies larger kurtosis and sparser temporal transitions of the generative factors (special cases are Gaussian, $\alpha = 2$, and Laplacian, $\alpha = 1$). Critically, for our proof we assume $\alpha < 2$ to ensure that temporal transitions are sparse. The novelty of our approach lies in our explicit modeling of sparse transitions that cover the statistics of natural data, which results in a stronger identifiability proof than previously achieved (see Appendix F.1.1 for a more detailed comparison with Hyvärinen and Morioka, 2017; Khemakhem et al., 2020a).

3.2 IDENTIFIABILITY PROOF

Theorem 1 *For a ground-truth $(g^*, \lambda^*, \alpha^*)$ and a learned (g, λ, α) model as defined in Eq. (2), if the functions g^* and g are injective and differentiable almost everywhere, $\lambda^* = \lambda$, $\alpha^* = \alpha < 2$ (i.e. there is no model misspecification) and the distributions of pairs of images generated from the priors $\mathbf{z}^* \sim p^*(\mathbf{z})$ and $\mathbf{z} \sim p(\mathbf{z})$ generated as $(g^*(\mathbf{z}^*_{t-1}), g^*(\mathbf{z}^*_t))$ and $(g(\mathbf{z}_{t-1}), g(\mathbf{z}_t))$, respectively, are matched almost everywhere, then $g = g^* \circ \sigma$, where σ is composed of a permutation and sign flips.*

The formal proof is provided in Appendix A.1. Similar to linear ICA, but in the temporal domain, we have to assume that the transitions of generative factors across time be non-Gaussian. Specifically, if the temporal changes of ground-truth factors are sparse, then the only generator consistent with the observations is the ground-truth one (up to a permutation and sign flips). The main idea behind the proof is to represent g as $g^* \circ h$ and note that if h were not a permutation, then the distributions $((g^* \circ h)(\mathbf{z}_{t-1}), (g^* \circ h)(\mathbf{z}_t))$ and $(g^*(\mathbf{z}^*_{t-1}), g^*(\mathbf{z}^*_t))$ would not match, due to the injectivity of g^*. Whether or not these distributions are the same is equivalent to whether or not the distributions of pairs (z_{t-1}, z_t) and $(h(z_{t-1}), h(z_t))$ are the same. For these distributions to be the same, the function h must preserve the Gaussian marginal for the first time step as well as the joint distribution, implying that it must preserve both the vector lengths and distances in the latent space. As we argue in the extended proof, this can only be the case if h is a composition of permutations and sign flips.

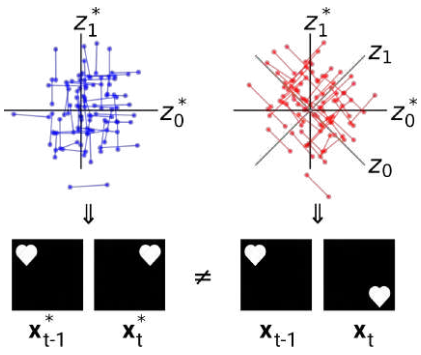


Figure 2: **Proof Intuition**. Latent representation and example generated image pairs for ground-truth (blue) and entangled (red) model. See text below for details.

Intuition Fig. 2 illustrates, by contradiction, why the model defined in Eq. (2) is identifiable. We consider temporal pairs of latents represented by connected points. A sparse transition prior encourages axis-alignment, as can be seen from the Laplace transition prior in the third image of Fig. 3.

[2]For a stationary stochastic process, $p(\mathbf{z}_{t-1})$ represents the instantaneous marginal distribution and $p(\mathbf{z}_t|\mathbf{z}_{t-1})$ the transition distribution. In case of an autoregressive process with non-Gaussian innovations with finite variance, it follows from the central limit theorem that the marginal distribution converges to a Gaussian in the limit of large λ.

This results in lines that are parallel with the axes in both the ground truth (left, blue, $\mathbf{z}^*$) and learned model (right, red, $\mathbf{z}$). In this example, z_0^* corresponds to horizontal position, while z_1^* corresponds to vertical position. The learned model must satisfy two criteria: (1) the latent factors should match the sparse prior (axis-aligned) and (2) the generated image pairs should match the ground-truth image pairs. If the learned latent factors were mismatched, for example by rotation, then the image pair distributions would not be matched. In this example, the ground truth model would produce image pairs with typically vertical or horizontal transitions, while the learned model pairs result in mostly diagonal transitions. Thus, the learned model cannot satisfy both criteria without aligning the latent axes with the ground-truth axes.

3.3 Slow Variational Autoencoder

In order to validate our proof, we must choose a probabilistic latent variable model for estimating the data density. We chose to build upon the framework of VAEs because of their efficiency in estimating a variational approximation to the ground truth posterior of a deep latent variable model (Kingma and Welling, 2013). We will refer to this model as *SlowVAE*. In Appendix B we note shortcomings of such an approach and test an alternative flow-based model.

The standard VAE objective assumes *i.i.d.* data and a standard normal prior with diagonal covariance on the learned latent representations $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To extend this to sequences, we assume the same functional form for our model prior as in Eq. (1) and Eq. (2). The posterior of our model is independent across time steps. Specifically,

$$q(\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{x}_t, \mathbf{x}_{t-1}) = q(\mathbf{z}_t|\mathbf{x}_t)\, q(\mathbf{z}_{t-1}|\mathbf{x}_{t-1}), \quad q(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{d} \mathcal{N}(\mu_i(\mathbf{x}), \sigma_i^2(\mathbf{x})), \tag{3}$$

where $\mu_i(\mathbf{x})$ and $\sigma_i^2(\mathbf{x})$ are the input-dependent mean and variance of our model's posterior. We visualize this combination of priors and posteriors in Fig. 3. For a given pair of inputs $(\mathbf{x}_t, \mathbf{x}_{t-1})$, the full evidence lower bound (ELBO, which we derive in Appendix A.2) can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{x}_t, \mathbf{x}_{t-1}) = E_{q(\mathbf{z}_t, \mathbf{z}_{t-1}|\mathbf{x}_t, \mathbf{x}_{t-1})}[\log p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{z}_t, \mathbf{z}_{t-1})] - D_{KL}(q(\mathbf{z}_{t-1}|\mathbf{x}_{t-1})|p(\mathbf{z}_{t-1})) \\ - \gamma\, E_{q(\mathbf{z}_{t-1}|\mathbf{x}_{t-1})}[D_{KL}(q(\mathbf{z}_t|\mathbf{x}_t)|p(\mathbf{z}_t|\mathbf{z}_{t-1}))],\end{aligned} \tag{4}$$

where γ is a regularization term for the sparsity prior, analogous to β in β-VAEs (Higgins et al., 2017) (technically, Eq. 4 is only an ELBO with $\gamma \leq 1$). The first term on the right-hand side is the log-likelihood (i.e. the negative reconstruction error, with $p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{z}_t, \mathbf{z}_{t-1})$ parameterized by the decoder of the VAE), the second term is the KL to a normal prior as in the standard VAE and the last term is an expectation of the KL between the posterior at time step t and the conditional prior $p(\mathbf{z}_t|\mathbf{z}_{t-1})$. The expectation in the last term is taken over samples from the posterior at the previous time step $q(\mathbf{z}_{t-1}|\mathbf{x}_{t-1})$. We observed empirically that taking the mean, $\mu(\mathbf{x}_{t-1})$, as a single sample produces good results, analogous to the log-likelihood that is typically evaluated at a single sample from the posterior (see Blei et al. (2017) for context).

In practice, we need to choose α, λ, and γ. For the latter two, we can perform a random search for hyperparameters, as we discuss below. For the former, any $\alpha < 2$ would break the general rotation symmetry by having an optimum for axis-aligned representations, which theorem 1 includes as a requirement for identifiability. As can be seen in Figs. 1 and 11, $\alpha \approx 0.5$ provides the best fit to the ground-truth marginals. However, we used $\alpha = 1$ as a parsimonious choice for SlowVAE, since the Laplace is a well-understood distribution that allows us to derive a simple closed-form solution for the ELBO in Eq. 4, which we derive in Appendix A.2.

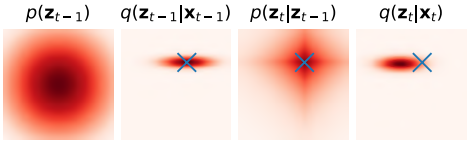


Figure 3: **SlowVAE illustration.** The prior and posterior for a two-dimensional latent space. Left to right: Normal prior for $t-1$, posterior for $t-1$, conditional Laplace prior for t, and posterior for t. The blue cross in the right three plots indicates the mean of the posterior for $t-1$.

3.4 Towards an Approximate Theory of Disentanglement

A number of our theoretical assumptions are violated in practice: After non-convex optimization, on a finite data sample, the distributions $p(\mathbf{x}_t, \mathbf{x}_{t-1})$ and $p^*(\mathbf{x}_t, \mathbf{x}_{t-1})$ are probably not perfectly

matched. In addition, the model assumptions on $p(\mathbf{z}_t, \mathbf{z}_{t-1})$ likely do not fully match the distribution of the ground truth factors. For example, the model may be misspecified such that $\alpha \neq \alpha^*$ or $\lambda \neq \lambda^*$, or the chosen family of distributions may be incorrect altogether. In the following section we will present results on several datasets where the marginal distributions $p(\mathbf{z}_{t-1})$ are drawn from a Uniform (not Normal) distribution, and some of them are over unordered sets (categories) or bounded periodic spaces (rotation). Also, in practice the model latent space is usually chosen to have more dimensions than the ground truth generative model. On real data, factors of variation may be dependent (Träuble et al., 2020; Yang et al., 2020). We show this is the case on YouTube-VOS and KITTI-MOTS in Appendix G and we provide evidence that breaking these dependencies has no clear consequence on disentanglement in Appendix F.2. A more formal treatment of dependence is done by Khemakhem et al. (2020b) who relax the independence assumption of ICA to Independently Modulated Components Analysis (IMCA) and introduce a family of conditional energy-based models that are identifiable up to simple transformations. Furthermore, the hypothesis class $\mathcal{G}$ of learnable functions in the VAE architecture may not contain the invertible ground truth generator $g^* \notin \mathcal{G}$, if it exists at all (e.g. occlusions may already lead to non-invertibility). Despite these violations, we consider it a strength of our method that the practical implementation still achieves improved disentanglement over previous approaches. However, we note understanding the impact of these violations as an important focus area for continued progress towards developing a practical yet theoretically supported method for disentanglement on natural scenes.

4 Datasets with Natural Transitions

While the standard datasets compiled by DisLib are an important step towards real-world applications, they still assume the data is *i.i.d.*. As described in section 2, Locatello et al. (2020) proposed uniformly sampling the number of factors to be changed, $k = \text{Rnd}$, and changing said factors by uniformly sampling over the possible set of values. What we refer to as "UNI" is a dataset variant modeled after the described scheme (Locatello et al., 2020) (further details in Appendix D). Considering our natural data analysis presented in Figure 1, such transitions are certainly unnatural. Given the current state of evaluation, we provide a set of incrementally more natural datasets which are otherwise comparable to existing work. We propose that said datasets should be included in the standard benchmark suite to provide a step towards disentanglement in natural data.

(1) Laplace Transitions (LAP) is a procedure for constructing image pairs from DisLib datasets by sampling from a sparse conditional distribution. For each ground-truth factor, the first value in the pair is chosen *i.i.d.* from the dataset and the second is chosen by weighting nearby factor values using Laplace distributed probabilities. LAP is a step towards natural data that closely resembles previous extensions of DisLib datasets to the time domain, but in a way that matches the marginal distribution of natural transitions (see Appendix D.2 for more details).
(2) Natural Sprites consists of pairs of rendered sprite images with generative factors sampled from real YouTube-VOS transitions. For a given image pair, the position and scale of the sprites are set using measured values from adjacent time points in YouTube-VOS. The sprite shapes and orientations are simple, like dSprites, and are fixed for a given pair. While fixing shape follows the natural transitions of objects, it is unclear how to accurately estimate object orientation from the masks, and thus we fixed the factor to avoid introducing artificial transitions. We additionally consider a version that is discretized to the same number of object states as dSprites, which i) allows us to use the standard DisLib evaluation metrics and ii) helps isolate the effect of including natural transitions from the effect of increasing data complexity (see Appendix D.4 for more details).
(3) KITTI Masks is composed of pedestrian segmentation masks from the autonomous driving vision benchmark KITTI-MOTS, thus with natural shapes and continuous natural transitions in all underlying factors. We consider adjacent frames which correspond to mean$(\Delta t) = 0.05s$ in physical time (we report the mean because of variable sampling rates in the original data); as well as frames with a larger temporal gap of mean$(\Delta t) = 0.15s$, which corresponds to samples of pairs that are at most 5 frames apart. We show in Appendix G.3 that SlowVAE disentanglement performance increases and then plateaus as we continue to increase mean(Δt).

In summary, we construct datasets with (1) imposed sparse transitions, (2) augmented with natural continuous generative factors using measurements from unstructured natural videos, as well as (3) data from unstructured natural videos themselves, but provided as segmentation masks to ensure visual complexity is manageable for current methods. For the provided datasets, the object categories

Model	Data	BetaVAE	FactorVAE	MIG	MCC	DCI	Modularity	SAP
PCL	dSprites (Uniform)	80.1 (0.4)	62.1 (0.9)	16.0 (7.4)	41.6 (1.5)	42.4 (1.2)	**99.7** (0.6)	6.0 (2.7)
Ada-GVAE	dSprites (Uniform)	88.0 (2.7)	73.1 (3.9)	17.3 (4.7)	46.0 (4.8)	32.3 (4.6)	93.3 (1.8)	6.6 (2.0)
SlowVAE	dSprites (Uniform)	87.0 (5.1)	75.2 (11.1)	**28.3** (11.5)	**58.8** (8.9)	47.7 (8.5)	86.9 (2.8)	4.4 (2.0)
PCL	dSprites (Laplace)	99.9 (0.1)	94.7 (3.1)	19.2 (3.1)	67.9 (3.3)	52.0 (3.5)	93.2 (0.9)	8.1 (1.6)
Ada-GVAE	dSprites (Laplace)	91.4 (1.6)	83.0 (5.9)	21.8 (4.9)	56.9 (4.2)	39.0 (4.2)	87.6 (1.8)	7.2 (0.3)
SlowVAE	dSprites (Laplace)	100.0 (0.0)	97.5 (3.0)	**29.5** (9.3)	69.8 (2.3)	**65.4** (3.6)	**96.5** (1.6)	8.1 (3.0)
PCL	Natural (Discrete)	82.4 (6.7)	68.3 (8.0)	7.8 (2.8)	50.2 (4.2)	14.3 (3.0)	88.9 (3.1)	2.5 (1.1)
Ada-GVAE	Natural (Discrete)	83.4 (1.1)	74.8 (4.4)	14.5 (3.2)	51.6 (2.5)	21.8 (2.9)	87.8 (2.5)	5.3 (1.4)
SlowVAE	Natural (Discrete)	82.6 (2.2)	76.2 (4.8)	11.7 (5.0)	52.6 (4.1)	18.9 (5.5)	88.1 (3.6)	4.4 (2.3)

Table 1: Mean and standard deviation (s.d.) metric scores across 10 random seeds. PCL is a scaled-up implementation of the method described by Hyvärinen and Morioka (2017), leveraging the encoding architecture and training hyperparameters specified in appendix E. Ada-GVAE is the leading method proposed by Locatello et al. (2020). Bold indicates statistical significance above the next highest score (independent T-test, $p < 0.05$). Red indicates statistical significance below the next lowest score. Results for additional datasets and models are in Table 2 and Appendix G.

never change across transitions – reflecting natural object permanence. Finally, as (2) and (3) use factor transitions measured from natural videos, they exhibit any natural statistical structure present for those factors, such as natural dependencies (further discussion is in Appendix F.2).

5 Experiments

5.1 Empirical Studies

We evaluate models using the DisLib implementation for the following supervised metrics: BetaVAE (Higgins et al., 2017); FactorVAE (Kim and Mnih, 2018); Mutual Information Gap (MIG; Chen et al., 2018); Disentanglement, Compactness, and Informativeness (DCI / Disentanglement; Eastwood and Williams, 2018); Modularity (Ridgeway and Mozer, 2018); and Separated Attribute Predictability (SAP; Kumar et al., 2018) (see Appendix C for metric details). None of the DisLib metrics support ground-truth labels with continuous variation, which is required for evaluation on the continuous Natural Sprites and KITTI Masks datasets. To reconcile this, we measure the Mean Correlation Coefficient (MCC), a standard metric in the ICA literature that is applicable to continuous variables. We report mean and standard deviation across 10 random seeds.

In order to select the conditional prior regularization and the prior rate in an unsupervised manner, we perform a random search over $\gamma \in [1, 16]$ and $\lambda \in [1, 10]$ and compute the recently proposed unsupervised disentanglement ranking (UDR) scores (Duan et al., 2020). We notice that the optimal values are close to $\gamma = 10$ and $\lambda = 6$ on most datasets, and thus use these values for all experiments. We leave finding optimal values for specific datasets to future work, but note that it is a strong advantage of our approach that it works well with the same model specification across 13 datasets (counting LAP and UNI for DisLib and optional discretization for Natural Sprites), addressing a concern posed in (Locatello et al., 2018). Additional details on model selection and training can be found in Appendix E. Although we train on image pairs, our model does not need paired data points at test time. For all visualizations, we pick the models with the highest average score across the DisLib metrics.

To compare our model fairly against other methods that also take image pairs as inputs, we also present performance for Permutation-Contrastive Learning from nonlinear ICA (PCL, Hyvärinen and Morioka, 2017) and Ada-GVAE, the leading method in the study by (Locatello et al., 2020). We scaled up the implementation of PCL for evaluation on our high-dimensional pixel inputs, and note this method does not have any hyperparameters. For Ada-GVAE, following the paper's recommendations, we select β (per dataset) using the considered parameter set $[1, 2, 4, 6, 8, 16]$, and use the reconstruction loss as the unsupervised model selection criterion (Locatello et al., 2020).

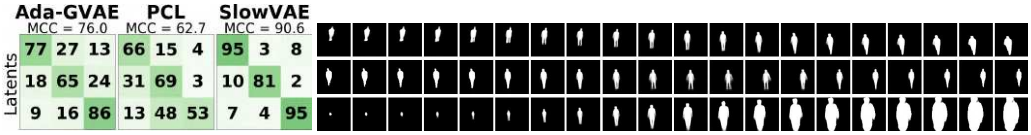


Figure 4: **KITTI Masks** (mean$(\Delta t) = 0.15s$)**.** (Left) MCC correlation matrix of the top 3 latents corresponding to y-position, x-position and scale. (Right) Images produced by varying the SlowVAE latent unit that corresponds to the corresponding row in the MCC matrix.

5.2 Results on DisLib and New Benchmarks

In Table 1 we demonstrate favorable performance compared to PCL and Ada-GVAE across all applicable metrics for discrete ground-truth variable datasets. The relative improvement on UNI is particularly surprising given the drastic mismatch between UNI and SlowVAE's assumptions. In Appendix G, we report results for the remaining DisLib datasets, where the observed dSprites results largely transfer. We also outperform PCL with a (flow-based) exact likelihood implementation of our slow transition prior in Appendix F.1.1. In Appendix F.3, we show that a model with an L_2 transition ($\alpha = 2$) prior performs much worse, supporting our theoretical prediction.

On the KITTI Masks dataset, one source of variation in the data is the average temporal separation within pairs of images mean(Δt). We present two settings (mean$(\Delta t) = 0.05s$, mean$(\Delta t) = 0.15s$) and observe a comparative increase in MCC for the latter (Table 2). Namely, the increase in performance for larger time gap is more pronounced with SlowVAE than the baselines, resulting in a statistically significant MCC gain. We provide details on the settings and ablate over the mean(Δt) parameter in Appendix G.3, where we observe a positive trend between mean(Δt) and MCC (reflecting Table 2, in Oord et al., 2018). Finally, we also verify that the transition distributions remain sparse despite the increase in this parameter (Appendix G.3). In Fig. 4, we can see that SlowVAE has learned latent dimensions which have correspondence with the estimated ground truth factors of x/y-position and scale.

Model	Data	MCC
PCL	Natural (Continuous)	51.7 (3.0)
Ada-GVAE	Natural (Continuous)	48.4 (4.8)
SlowVAE	Natural (Continuous)	49.1 (4.0)
PCL	Kitti (mean$(\Delta t) = 0.05s$)	52.6 (5.1)
Ada-GVAE	Kitti (mean$(\Delta t) = 0.05s$)	62.6 (7.5)
SlowVAE	Kitti (mean$(\Delta t) = 0.05s$)	66.1 (4.5)
PCL	Kitti (mean$(\Delta t) = 0.15s$)	58.5 (3.3)
Ada-GVAE	Kitti (mean$(\Delta t) = 0.15s$)	67.6 (6.7)
SlowVAE	Kitti (mean$(\Delta t) = 0.15s$)	**79.6** (5.8)

Table 2: Continuous ground-truth variable datasets. See Table 1 for details.

Locatello et al. (2018) showed that all *i.i.d.* models performed similarly across the DisLib datasets and metrics when testing was carefully controlled. However, in Fig. 5 we observe that the different modeling assumptions result in differences in representation quality. To construct the visuals, we first compute the sorted correlation matrix between the latents (rows) and generative factors (columns), which we visualize as a correlation matrices. The matrices are sorted via linear sum assignment such that each ground-truth factor is non-greedily associated with the latent variable with highest correlation (Hyvärinen and Morioka, 2016). Below the matrices are scatter plots that reveal the decodability of the assigned latent factors. In each scatter plot, the horizontal axis indicates the ground truth value, the vertical axis indicates the corresponding latent value, and the colors indicate object shape. The models displayed are those with the maximum average score across evaluated metrics.

The latent space visualizations use the known ground-truth factors to aid in understanding how each factor is encoded in a way that is more informative than exclusively visualizing latent traversals or embeddings of pairs of latent units (Cheung et al., 2014; Chen et al., 2016; Szabó et al., 2017; Ma et al., 2018). For example, in the third row, we observe that several models have a sinusoidal variation with frequencies $\sim \omega, 2\omega$, and 4ω, which correspond to the three distinct rotational symmetries of the shapes: heart, ellipse and square. This directly impacts MCC performance (third row in the MCC matrix), which measures rank correlation between the matching latent factor (an angular variable) and the ground truth, which encodes the angles with monotonically increasing indices. Furthermore, the square has a four-fold rotational symmetry and repeats after 90°, but it is represented in a full 360° rotation in the DisLib ground truth encoding format, resulting in different ground truth labels for identical input images.

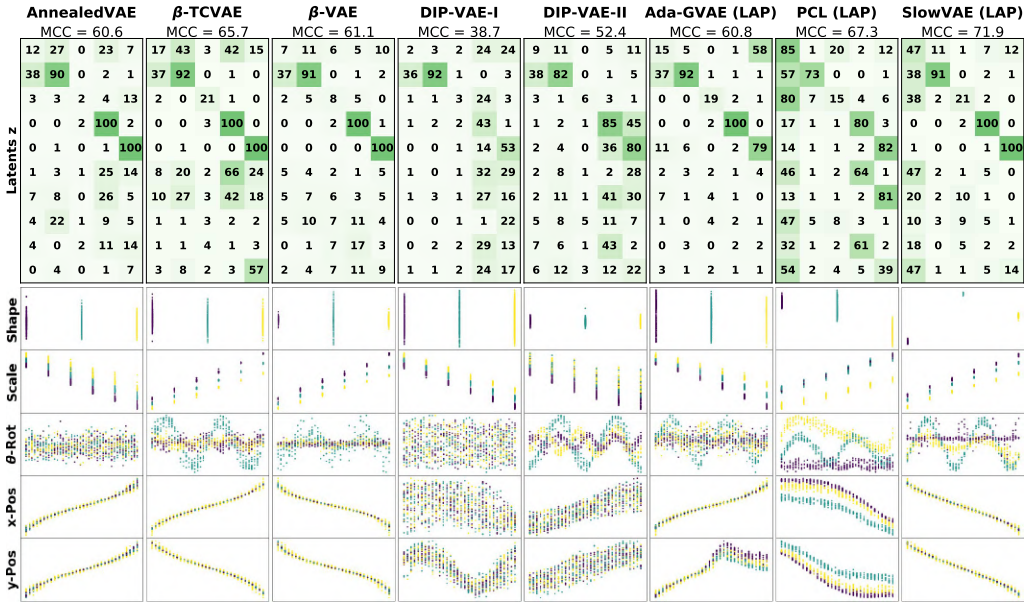


Figure 5: **DSprites Latent Representations:** (Top) shows absolute MCC between generative and model factors (rows are rearranged for maximal correlation on the main diagonal). The columns correspond to generative factors (shape, scale, rotation, x/y-position) and the values correspond to percent correlation. A more diagonal structure in the upper half corresponds to a better one-to-one mapping between generative and latent factors. (Bottom) shows individual latent dimensions (y-axis) over the matched generative factors (x-axis). Colors encode shapes: heart/yellow, ellipse/turquoise, and square/purple.

A similar observation can be made with respect to the categorical factors, which are also represented as ordinal ground truth variables. For example, the PCL correlation score (top left element in the PCL MCC matrix) is quite high, while the corresponding shape correlation score for SlowVAE is quite low. However, if we consider the shape scatter plots, we clearly see that SlowVAE separates the three shapes more distinctively than PCL, only in an order that differs from the ground truth. One solution is to modify MCC to report the maximum correlation over all permutations of the ground truth assignments, although brute force methods for this would scale poorly with the number of categories. We also note that datasets where we see small performance differences among models (e.g., Cars3D) have significantly more discrete categories (e.g., 183) than the other datasets ($3 - 6$). This could also explain why all models considered in Table 1 and 2 perform comparably on the Natural Sprites datasets, where unlike KITTI Masks the ground truth evaluation includes categorical and angular variables. We note that properly evaluating disentanglement is an ongoing area of research (Duan et al., 2020), with notable preliminary results in recent work (Higgins et al., 2018; Bouchacourt et al., 2021; Tonnaer et al., 2020).

6 Conclusion

We provide evidence to support the hypothesis that natural scenes exhibit highly sparse marginal transition probabilities. Leveraging this finding, we contribute a novel nonlinear ICA framework that is provably identifiable up to permutations and sign-flips — a stronger result than has been achieved previously. With the SlowVAE model we provide a parsimonious implementation that is inspired by a long history of learning visual representations from temporal data (Sutton, 1988; Hinton, 1990; Földiák, 1991). We apply this model to current metric-based disentanglement benchmarks to demonstrate that it outperforms existing approaches (Locatello et al., 2020; Hyvärinen and Morioka, 2017) on aggregate without any tuning of hyperparameters to individual datasets. We also provide novel video dataset benchmarks to guide disentanglement research towards more natural domains.

We observe that these datasets have complex dependencies that our theory will have to be extended to account for, although we demonstrate with empirical comparisons the efficacy of our approach. In addition to Natural Sprites and KITTI Masks, we suggest that YouTube-VOS will be valuable as a large-scale dataset that is unconstrained by object type and scenario for more advanced models. Variance in such categorical factors is problematic for evaluation due to the cited drawbacks of existing quantitative metrics, which should be addressed in tandem with scaling to natural data. Taken together, our dataset and model proposals set the stage for utilizing knowledge of natural scene statistics to advance unsupervised disentangled representation learning.

In our experiments we see that approximate identification as measured by the different disentanglement metrics increases despite violations of theoretical assumptions, which is in line with prior studies (Shu et al., 2019; Khemakhem et al., 2020a; Locatello et al., 2020). Nevertheless, future work should address gaining a better understanding of the theoretical and empirical consequences of such model misspecifications, in order to make the theory of disentanglement more predictive about empirically found solutions.

ACKNOWLEDGEMENTS

The authors would like to thank Francesco Locatello for valuable discussions and providing numerical results to facilitate our experimental comparisons. Additionally, we thank Luigi Gresele, Matthias Tangemann, Roland Zimmermann, Robert Geirhos, Matthias Kümmerer, Cornelius Schröder, Charles Frye, and Sarah Master for helpful feedback in preparing the manuscript. Finally, the authors would like to thank Johannes Ballé, Jon Shlens and Eero Simoncelli for early discussions related to the ideas developed in this paper.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the priority program 1835 under grant BR2321/5-2 and by SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms (TP3), project number: 276693517. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting LS and YS. DP was supported by the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A). IU, WB, and MB are supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

The authors declare no conflicts of interests.

BROADER IMPACT

Representation learning is at the heart of model building for cognition. Our specific contribution is focused on core methods for modeling natural videos and the datasets used are more simplistic than real-world examples. However, foundational research on unsupervised representation learning has potentially large impact on AI for advancing the power of self-learning systems.

The broader field of representation learning has a large number of focused research directions that span machine learning and computational neuroscience. As such, the application space for this work is vast. For example, applications in unsupervised analysis of complicated and unintuitive data, such as medical imaging and gene expression information, have great potential to solve fundamental problems in health sciences. A future iteration of our disentangling approach could be used to encode such complicated data into a lower-dimensional and more understandable space that might reveal important factors of variation to medical researchers. Another important and complex modeling space that could potentially be improved by this line of research is in environmental sciences and combating global climate change.

Nonetheless, we acknowledge that any machine learning method can be used for nefarious purposes, which can be mitigated via effective, scientifically informed communication, outreach, and policy direction. We unconditionally denounce the use of derivatives of our work for weaponized or wartime applications. Additionally, due to the lack of interpretability generally found in modern deep learning approaches, it is possible for practitioners to inadvertently introduce harmful biases or errors in machine learning applications. Although we certainly do not solve this problem, our focus on providing identifiable solutions to representation learning is likely beneficial for both interpretability and fairness in machine learning.

REFERENCES

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pages 50–59, 2018.

Jonathan T Barron and Jitendra Malik. Shape, albedo, and illumination from a single image of an unknown object. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 334–341. IEEE, 2012.

Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, Apr 2017. ISSN 1537-274X. doi: 10.1080/01621459.2017.1285773. URL http://dx.doi.org/10.1080/01621459.2017.1285773.

Diane Bouchacourt, Mark Ibrahim, and Stéphane Deny. Addressing the topological defects of disentanglement via distributed operators, 2021.

Samuel R. Bowman, L. Vilnis, Oriol Vinyals, Andrew M. Dai, R. Józefowicz, and S. Bengio. Generating sentences from a continuous space. In *CoNLL*, 2016.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Christpher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

Charles F Cadieu and Bruno A Olshausen. Learning intermediate-level representations of form and motion from natural movies. *Neural Computation*, 24(4):827–866, 2012.

J-F Cardoso. Source separation using higher order moments. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 2109–2112. IEEE, 1989.

Tian Qi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, page 1436–1445, 2019.

Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423, 2017.

Adji B. Dieng, Yoon Kim, Alexander M. Rush, and D. Blei. Avoiding latent variable collapse with generative skip models. *ArXiv*, abs/1807.04863, 2019.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *ArXiv*, abs/1410.8516, 2017a.

Laurent Dinh, Jascha Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2017b.

Sunny Duan, Loic Matthey, Andre Saraiva, Nick Watters, Christopher Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *In International Conference on Learning Representations*, 2018.

Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3(2):194–200, 1991.

Lijian Gao, Qirong Mao, Ming Dong, Yu Jing, and Ratna Chinnam. On learning disentangled representation for acoustic event detection. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2006–2014, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *Advances in Neural Information Processing Systems*, pages 15714–15725, 2019.

Will Grathwohl and Aaron Wilson. Disentangling space and time in video with hierarchical variational auto-encoders. *arXiv preprint arXiv:1612.04440*, 2016.

Klaus Greff, Raphael Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450*, 2019.

Luigi Gresele, Giancarlo Fissore, Adrian Javaloy, Bernhard Scholkopf, and Aapo Hyvarinen. Relative gradient optimization of the jacobian term in unsupervised deep learning. *ArXiv*, abs/2006.15090, 2020.

Wakako Hashimoto. Quadratic forms in natural images. *Network: Computation in Neural Systems*, 14(4): 765–788, 2003.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. *arXiv preprint arXiv:1901.05534*, 2019.

Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, 22(6):984–991, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2(5):6, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Geoffrey E Hinton. Connectionist learning procedures. In *Machine learning*, page 208. Elsevier, 1990.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Proceedings of Machine Learning Research*, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Jarmo Hurri, and Jaakko Väyrynen. Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *JOSA A*, 20(7):1237–1252, 2003.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

Christian Jutten and Jeanny Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020a.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33, 2020b.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pages 4743–4751, 2016.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pages 14611–14624, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, pages 9408–9418, 2019.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. In *Advances in neural information processing systems*, pages 6551–6562, 2019.

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4402–4412, 2019.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Stanisław Mazur and Stanisław Ulam. Sur les transformations isométriques d'espaces vectoriels normés. *CR Acad. Sci. Paris*, 194(946-948):116, 1932.

Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, March 2016.

Graeme Mitchison. Removing time variation with the anti-hebbian differential synapse. *Neural Computation*, 3 (3):312–320, 1991.

Hossein Mobahi, Ronan Collobert, and Jason Weston. Deep learning from temporal coherence in video. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 737–744, 2009.

Hiroshi Morioka. Time-contrastive learning (tcl), 2018. URL https://github.com/hirosm/TCL.

Bruno A Olshausen. Learning sparse, overcomplete representations of time-varying natural images. In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, volume 1, pages I–41. IEEE, 2003.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Edouard Pineau, S. Razakarivony, and T. Bonald. Time series source separation with slow flows. *ArXiv*, abs/2007.10182, 2020.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in neural information processing systems*, pages 1252–1260, 2015.

Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Fabian Sinz, Sebastian Gerwinn, and Matthias Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820, 2009.

Jascha Sohl-Dickstein, Ching Ming Wang, and Bruno A Olshausen. An unsupervised algorithm for learning lie group transformations. *arXiv preprint arXiv:1001.1027*, 2010.

Peter Sorrenson, Carsten Rother, and Ulrich Kothe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *ArXiv*, abs/2001.04872, 2017.

Mikhail Fedorovich Subbotin. On the law of frequency of error. *Mat. Sb.*, 31(2):296–301, 1923.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

Attila Szabó, Qiyang Hu, Tiziano Portenier, Matthias Zwicker, and Paolo Favaro. Challenges in disentangling independent factors of variation. *arXiv preprint arXiv:1711.02245*, 2017.

Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

Esteban G Tabak, Eric Vanden-Eijnden, et al. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010.

Loek Tonnaer, Luis A. Pérez Rey, Vlado Menkovski, Mike Holenderski, and Jacobus W. Portegies. Quantifying and learning disentangled representations with limited supervision, 2020.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020.

Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. *arXiv preprint arXiv:1912.02783*, 2019.

Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

Rishi Veerapaneni, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua B. Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based reinforcement learning. *arXiv preprint arXiv:1910.12827*, 2019.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Nicholas Watters, Loic Matthey, Sebastian Borgeaud, Rishabh Kabra, and Alexander Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment, 2019. URL `https://github.com/deepmind/spriteworld/`.

Marissa A. Weis, Kashyap Chitta, Yash Sharma, Wieland Brendel, Matthias Bethge, Andreas Geiger, and Alexander S. Ecker. Unmasking the inductive biases of unsupervised object representations for video sequences. *arXiv preprint arXiv:2006.07034*, 2020.

Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Markus Wulfmeier, Arunkumar Byravan, Tim Hertweck, Irina Higgins, Ankush Gupta, Tejas Kulkarni, Malcolm Reynolds, Denis Teplyashin, Roland Hafner, Thomas Lampe, and Martin Riedmiller. Representation matters: Improving perception and exploration for robotics. *arXiv preprint arXiv:2011.01758*, 2020.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *arXiv preprint arXiv:1905.04804*, 2019.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science Advances*, 6(10):eaax5979, 2020.

Will Zou, Shenghuo Zhu, Kai Yu, and Andrew Y Ng. Deep learning of invariant features via simulated fixations in video. In *Advances in neural information processing systems*, pages 3203–3211, 2012.

APPENDIX

A FORMAL METHODS

Function / variable	Description
g	Generator
α	Prior shape
λ	Prior rate
$p(\mathbf{z})$	Prior
$\mathbf{z} \sim p(\mathbf{z})$	Latent variables
$\mathbf{x} = g(\mathbf{z})$	Generated images
$q(\mathbf{z}\|\mathbf{x})$	Variational posterior

Table 3: Glossary of terms. We use a * (i.e. g^*) when necessary to highlight that we are referring to the ground truth model.

A.1 PROOF OF IDENTIFIABILITY

To study disentanglement, we assume that the generative factors $\mathbf{z} \in \mathbb{R}^D$ are mapped to images $\mathbf{x} \in \mathbb{R}^N$ (usually $D \ll N$, but see section B) by a nonlinear ground-truth generator $g^* : \mathbf{z} \mapsto \mathbf{x}$.

Theorem 1 *Let $(g^*, \lambda^*, \alpha^*)$ and (g, λ, α) respectively be ground-truth and learned generative models as defined in Eq.* (2)*. If the following conditions are satisfied:*

(i) The generators g^* and g are defined everywhere in the latent space. Moreover, they are injective and differentiable almost everywhere,

(ii) There is no model misspecification i.e. $\alpha = \alpha^*$ and $\lambda = \lambda^*$, so $\mathbf{z} \sim p(\mathbf{z}) = p^*(\mathbf{z})$,

(iii) Pairs of images are generated as $(\mathbf{x}^*_{t-1}, \mathbf{x}^*_t) = (g^*(\mathbf{z}_{t-1}), g^*(\mathbf{z}_t))$ and $(\mathbf{x}_{t-1}, \mathbf{x}_t) = (g(\mathbf{z}_{t-1}), g(\mathbf{z}_t))$,

(iv) The distributions of $(\mathbf{x}^*_{t-1}, \mathbf{x}^*_t)$ and $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ are the same (i.e. the corresponding densities are equal almost everywhere: $p^*(\mathbf{x}_{t-1}, \mathbf{x}_t) = p(\mathbf{x}_{t-1}, \mathbf{x}_t)$,

then $g = g^* \circ \sigma$, where σ is a composition of a permutation and sign flips.

Proof. Since $\mathbf{x} = g(\mathbf{z})$ can be written as $\mathbf{x} = (g^* \circ (g^*)^{-1} \circ g)(\mathbf{z})$, we can assume that $g = g^* \circ h$ for some function h on the latent space.

We first show that the function h is a bijection on the latent space. It is injective, since both g and g^* are injective. Because of continuity of h, if it were not surjective, there would be some neighborhood $\mathbf{U}_{\tilde{\mathbf{z}}}$ of $\tilde{\mathbf{z}}$ that would not have a pre-image under h. This would mean that images generated by g^* from $\mathbf{U}_{\tilde{\mathbf{z}}}$ would have zero density under the distribution of images generated by g (i.e. $p(g^*(\mathbf{U}_{\tilde{\mathbf{z}}})) = 0$). This density would be non-zero under the distribution of images directly generated by the ground-truth generator g^* (i.e. $p^*(g^*(\mathbf{U}_{\tilde{\mathbf{z}}})) \neq 0$), which contradicts the assumption that these distributions are equal. It follows that h is bijective.

In the next step, we show that the distribution of latent space pairs $(h(\mathbf{z}_{t-1}), h(\mathbf{z}_t))$ matches the latent space prior distribution (i.e. h preserves the prior distribution in the latent space). Indeed, using the assumption that the distributions of $(g^*(\mathbf{z}_{t-1}), g^*(\mathbf{z}_t))$ and $((g^* \circ h)(\mathbf{z}_{t-1}), (g^* \circ h)(\mathbf{z}_t))$ are the same, we can write the following equality using the change of variables formula:

$$\begin{aligned} p^*(\mathbf{x}_{t-1}, \mathbf{x}_t) &= p((g^*)^{-1}(\mathbf{x}_{t-1}), (g^*)^{-1}(\mathbf{x}_t)) \left| \det \left(\frac{\mathrm{d}(g^*)^{-1}}{\mathrm{d}(\mathbf{x}_{t-1}, \mathbf{x}_t)} \right) \right| \\ &= p_h((g^*)^{-1}(\mathbf{x}_{t-1}), (g^*)^{-1}(\mathbf{x}_t)) \left| \det \left(\frac{\mathrm{d}(g^*)^{-1}}{\mathrm{d}(\mathbf{x}_{t-1}, \mathbf{x}_t)} \right) \right| \\ &= p(\mathbf{x}_{t-1}, \mathbf{x}_t), \end{aligned} \tag{5}$$

where p and p_h are densities of $(\mathbf{z}_{t-1}, \mathbf{z}_t)$ and $(h(\mathbf{z}_{t-1}), h(\mathbf{z}_t))$. Since the determinants above cancel, these densities are equal at the pre-image of any pair of images $(\mathbf{x}_{t-1}, \mathbf{x}_t)$. Because g^* is defined

everywhere in the latent space, p and p_h are equal for any pair of latent space points. Applying the change of variables formula again, we obtain the following equation:

$$\begin{aligned} p(\mathbf{z}_{t-1}, \mathbf{z}_t) &= p(h^{-1}(\mathbf{z}_{t-1}), h^{-1}(\mathbf{z}_t)) \left|\det\left(\frac{\mathrm{d}h^{-1}}{\mathrm{d}(\mathbf{z}_{t-1}, \mathbf{z}_t)}\right)\right| \\ &= p(h^{-1}(\mathbf{z}_{t-1}))\, p(h^{-1}(\mathbf{z}_t) \mid h^{-1}(\mathbf{z}_{t-1})) \left|\det\left(\frac{\mathrm{d}h^{-1}(\mathbf{z}_{t-1})}{\mathrm{d}\mathbf{z}_{t-1}}\right)\right| \left|\det\left(\frac{\mathrm{d}h^{-1}(\mathbf{z}_t)}{\mathrm{d}\mathbf{z}_t}\right)\right| \\ &= p(\mathbf{z}_{t-1})\, p(\mathbf{z}_t \mid \mathbf{z}_{t-1}). \end{aligned} \tag{6}$$

Note that the probability measure p is the same before and after the change of variables, since we showed that the prior distribution in the latent space must be invariant under the function h. The same condition for the marginal $p(\mathbf{z}_{t-1})$ is as follows:

$$p(\mathbf{z}_{t-1}) = p(h^{-1}(\mathbf{z}_{t-1})) \left|\det\left(\frac{\mathrm{d}h^{-1}(\mathbf{z}_{t-1})}{\mathrm{d}\mathbf{z}_{t-1}}\right)\right|. \tag{7}$$

Solving for the determinant of the Jacobian in (7) and plugging it into (6), we obtain

$$p(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = p(h^{-1}(\mathbf{z}_t) \mid h^{-1}(\mathbf{z}_{t-1})) \frac{p(\mathbf{z}_t)}{p(h^{-1}(\mathbf{z}_t))}. \tag{8}$$

Taking logs of both sides, we arrive at the following equation:

$$A(||\mathbf{z}_t - \mathbf{z}_{t-1}||_\alpha^\alpha - ||h^{-1}(\mathbf{z}_t) - h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha) = B(||\mathbf{z}_t||_2^2 - ||h^{-1}(\mathbf{z}_t)||_2^2), \tag{9}$$

where A and B are the constants appearing in the exponentials in $p(\mathbf{z}_{t-1})$ and $p(\mathbf{z}_t \mid \mathbf{z}_{t-1})$. The logs of normalization constants cancel out.

For any $\mathbf{z}_t$ we can choose $\mathbf{z}_{t-1} = \mathbf{z}_t$ making the left hand side in (9) equal to zero. This implies that $||\mathbf{z}_t||_2^2 = ||h^{-1}(\mathbf{z}_t)||_2^2$ for any $\mathbf{z}_t$, i.e. function h^{-1} preserves the 2-norm. Moreover, the preservation of the 2-norm implies that $p(\mathbf{z}_{t-1}) = p(h^{-1}(\mathbf{z}_{t-1}))$ and therefore it follows from (7) that for any $\mathbf{z}$

$$\left|\det\left(\frac{\mathrm{d}h^{-1}(\mathbf{z})}{\mathrm{d}\mathbf{z}}\right)\right| = 1. \tag{10}$$

Thus, the left hand side of (9) can be re-written as

$$||\mathbf{z}_t - \mathbf{z}_{t-1}||_\alpha^\alpha - ||h^{-1}(\mathbf{z}_t) - h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha = 0. \tag{11}$$

This means that h^{-1} preserves the α-distances between points. Moreover, because h is bijective, the Mazur-Ulam theorem (Mazur and Ulam, 1932) tells us that h must be an affine transform.

In the next step, to prove that h must be a permutation and sign flip, let us choose an arbitrary point $\mathbf{z}_{t-1}$ and $\mathbf{z}_t = \mathbf{z}_{t-1} + \varepsilon\, \mathbf{e}_k = (z_{1,1}, \ldots, z_{1,k} + \varepsilon, \ldots, z_{1,D})$. Using (11) and performing a Taylor expansion around $\mathbf{z}_{t-1}$, we obtain the following:

$$\begin{aligned} \varepsilon^\alpha &= ||\mathbf{z}_t - \mathbf{z}_{t-1}||_\alpha^\alpha \\ &= ||h^{-1}(\mathbf{z}_{t-1} + \varepsilon\, \mathbf{e}_k) - h^{-1}(\mathbf{z}_{t-1})||_\alpha^\alpha \\ &= \left|\left|\varepsilon \cdot \left(\frac{\partial h_1^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}, \ldots, \frac{\partial h_D^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}\right) + O(\varepsilon^2)\right|\right|_\alpha^\alpha. \end{aligned} \tag{12}$$

The higher-order terms $O(\varepsilon^2)$ are zero since h is affine, therefore dividing both sides of the above equation by ε^α we find that

$$\left|\left|\left(\frac{\partial h_1^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}, \ldots, \frac{\partial h_D^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}\right)\right|\right|_\alpha^\alpha = 1. \tag{13}$$

The vectors of k-th partial derivatives of components of h^{-1} are columns of the Jacobian matrix $\left(\frac{\mathrm{d}h^{-1}(\mathbf{z})}{\mathrm{d}\mathbf{z}}\right)$. Using the fact that the determinant of that matrix is equal to one and applying Hadamard's inequality, we obtain that

$$\left|\det\left(\frac{\mathrm{d}h^{-1}(\mathbf{z})}{\mathrm{d}\mathbf{z}}\right)\right| = 1 \leq \prod_{k=1}^{D} \left|\left|\left(\frac{\partial h_1^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}, \ldots, \frac{\partial h_D^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}\right)\right|\right|_2. \tag{14}$$

Since $\alpha < 2$, for any vector $\mathbf{v}$ it holds that $||\mathbf{v}||_2 \leq ||\mathbf{v}||_\alpha$, with equality only if at most one component of $\mathbf{v}$ is non-zero. This inequality implies that both (13) and (14) hold at the same time if and only if

$$\left|\left|\left(\frac{\partial h_1^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}, \ldots, \frac{\partial h_D^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}\right)\right|\right|_2 = \left|\left|\left(\frac{\partial h_1^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}, \ldots, \frac{\partial h_D^{-1}(\mathbf{z}_{t-1})}{\partial z_{t-1,k}}\right)\right|\right|_\alpha = 1, \quad (15)$$

meaning that only one element of these vectors of k-th partial derivatives is non-zero, and it is equal to 1 or -1. Thus, the function h is a composition of a permutation and sign flips at every point. Potentially, this permutation might be input-dependent, but we argued above that h is affine, therefore the permutation must be the same for all points. □

A.2 Kullback Leibler Divergence of Slow Variational Autoencoder

The VAE learns a variational approximation to the true posterior by maximizing a lower bound on the log-likelihood of the empirical data distribution $\mathcal{D}$

$$\begin{aligned} &E_{\mathbf{x}_{t-1},\mathbf{x}_t\sim\mathcal{D}}[\log p(\mathbf{x}_{t-1},\mathbf{x}_t)] \geq \\ &\quad E_{\mathbf{x}_{t-1},\mathbf{x}_t\sim\mathcal{D}}[E_{q(\mathbf{z}_t,\mathbf{z}_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})}[\log p(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{z}_{t-1},\mathbf{z}_t) - \log q(\mathbf{z}_t,\mathbf{z}_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})]]. \end{aligned} \quad (16)$$

For this, we need to compute the Kullback-Leibler divergence (KL) between the posterior $q(\mathbf{z}_t,\mathbf{z}_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})$ and the prior $p(\mathbf{z}_t,\mathbf{z}_{t-1})$. Since all of these distributions are per design factorial, we will, for simplicity, derive the KL below for scalar variables (log-probabilities will simply have to be summed to obtain the full expression). Recall that the model prior and posterior factorize like

$$\begin{aligned} p(z_t,z_{t-1}) &= p(z_t|z_{t-1})\,p(z_{t-1}) \\ q(z_t,z_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1}) &= q(z_t|\mathbf{x}_t)\,q(z_{t-1}|\mathbf{x}_{t-1}). \end{aligned} \quad (17)$$

Then, given a pair of inputs $(\mathbf{x}_{t-1},\mathbf{x}_t)$, the KL can be written

$$\begin{aligned} &D_{KL}(q(z_t,z_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})|p(z_t,z_{t-1})) = E_{z_t,z_{t-1}\sim q(z_t,z_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})}\left[\log\frac{q(z_t|\mathbf{x}_t)\,q(z_{t-1}|\mathbf{x}_{t-1})}{p(z_t|z_{t-1})\,p(z_{t-1})}\right] \\ &= E_{z_{t-1}\sim q(z_{t-1}|\mathbf{x}_{t-1})}\left[\log\frac{q(z_{t-1}|\mathbf{x}_{t-1})}{p(z_{t-1})}\right] + E_{z_t,z_{t-1}\sim q(z_t,z_{t-1}|\mathbf{x}_t,\mathbf{x}_{t-1})}\left[\log\frac{q(z_t|\mathbf{x}_t)}{p(z_t|z_{t-1})}\right] \\ &= D_{KL}(q(z_{t-1}|\mathbf{x}_{t-1})|p(z_{t-1})) - H(q(z_t|\mathbf{x}_t)) + E_{z_{t-1}\sim q(z_{t-1}|\mathbf{x}_{t-1})}\left[H(q(z_t|\mathbf{x}_t),p(z_t|z_{t-1}))\right] \end{aligned} \quad (18)$$

Where we use the fact that KL divergences decompose like $D_{KL}(X,Y) = H(X,Y) - H(X)$ into (differential) cross-entropy $H(X,Y)$ and entropy $H(X)$. The first term of the last line in (18) is the same KL divergence as in the standard VAE, namely between a Gaussian distribution $q(z_{t-1}|\mathbf{x}_{t-1})$ with some $\mu(\mathbf{x}_{t-1})$ and $\sigma(\mathbf{x}_{t-1})$ and a standard Normal distribution $p(z_{t-1})$. The solution of the KL is given by $D_{KL}(q(z_{t-1}|\mathbf{x}_{t-1})|q(z_{t-1})) = -\log\sigma(\mathbf{x}_{t-1}) + \frac{1}{2}(\mu(\mathbf{x}_{t-1})^2 + \sigma(\mathbf{x}_{t-1})^2 - 1)$ (Bishop, 2006). The second term on the RHS, i.e. the entropy of a Gaussian is simply given by $H(q(z_t|\mathbf{x}_t)) = \log(\sigma(\mathbf{x}_t)\sqrt{2\pi e})$.

To compute the last term on the RHS, let us recall the Laplace form of the conditional prior

$$p(z_t|z_{t-1}) = \frac{\lambda}{2}\exp -\lambda|z_t - z_{t-1}|. \quad (19)$$

Thus the cross-entropy becomes

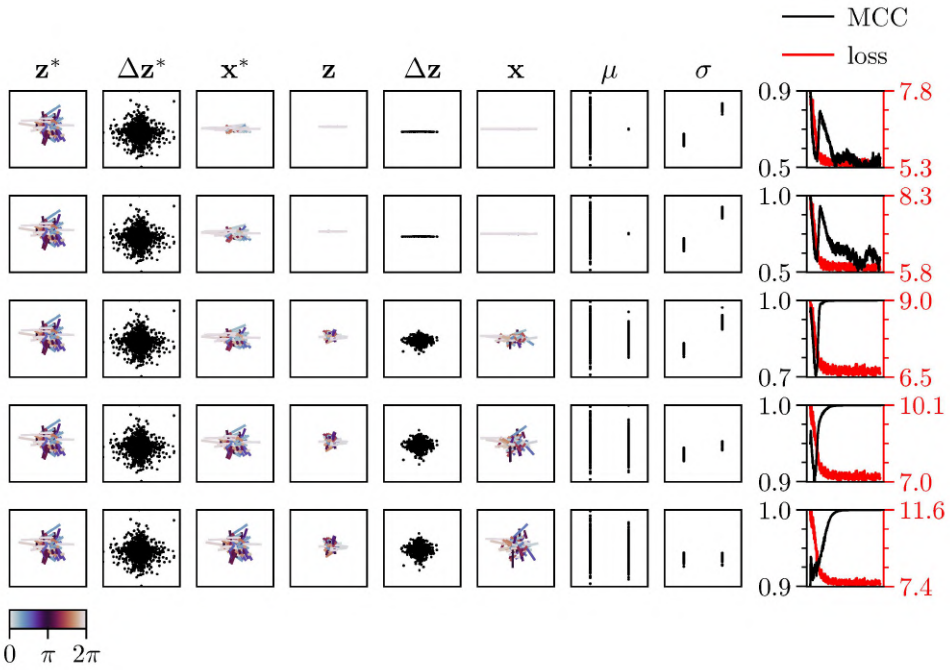
$$\begin{aligned} H(q(z_t|\mathbf{x}_t),p(z_t|z_{t-1})) &= -E_{z_t\sim q(z_t|\mathbf{x}_t)}[\log p(z_t|z_{t-1})] \\ &= -\log\left(\frac{\lambda}{2}\right) + \lambda E_{z_t\sim q(z_t|\mathbf{x}_t)}[|z_t - z_{t-1}|]. \end{aligned} \quad (20)$$

Now, if some random variable $X \sim \mathcal{N}(\mu,\, \sigma^2)$, then $Y = |X|$ follows a *folded normal distribution*, for which the mean is defined as

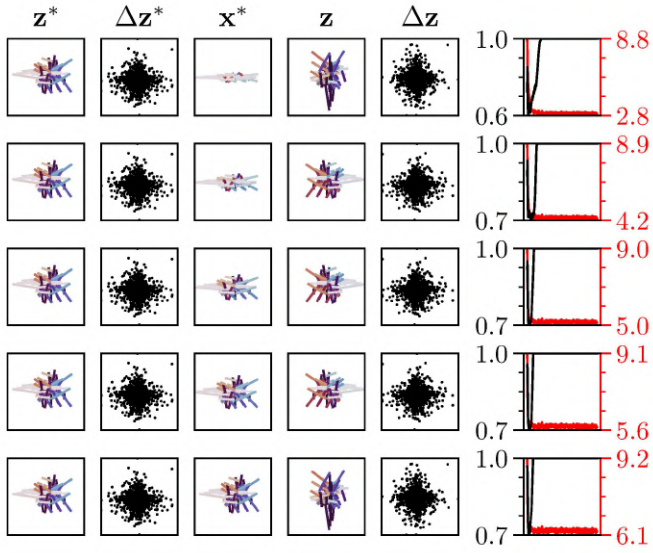
$$E[|x|] = \sigma\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) - \mu\left(1 - 2\,\Phi\left(\frac{\mu}{\sigma}\right)\right), \tag{21}$$

where Φ is the cumulative distribution function of a standard normal distribution (mean zero and variance one). Thus, denoting $\mu(\mathbf{x}_t)$ and $\sigma(\mathbf{x}_t)$ the mean and variance of $q(z_t|\mathbf{x}_t)$, and defining $\mu(\mathbf{x}_t, z_{t-1}) = \mu(\mathbf{x}_t) - z_{t-1}$, we can rewrite further

$$\begin{aligned} &H(q(z_t|\mathbf{x}_t), p(z_t|z_{t-1})) = \\ &-\log\left(\frac{\lambda}{2}\right) + \lambda\left(\sigma(\mathbf{x}_t)\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\mu(\mathbf{x}_t, z_{t-1})^2}{2\sigma(\mathbf{x}_t)^2}\right) - \mu(\mathbf{x}_t, z_{t-1})\left(1 - 2\,\Phi\left(\frac{\mu(\mathbf{x}_t, z_{t-1})}{\sigma(\mathbf{x}_t)}\right)\right)\right). \end{aligned} \tag{22}$$



(a) SlowVAE performance.



(b) SlowFlow performance.

Figure 6: **VAE failure modes.** Rows respectively indicate $\kappa = 0.2, 0.4, 0.6, 0.8, 1.0$ from Eq. (24). The left five columns show values for 100 randomly chosen examples, while the μ and σ columns show values for the full training set. Columns in the sets ($\mathbf{z}^*$, $\mathbf{z}$), ($\Delta\mathbf{z}^*$, $\Delta\mathbf{z}$), ($\mathbf{x}^*$, $\mathbf{x}$) all have the same (arbitrary) scale factors the axes. Lines indicate trajectories from time-point t to $t+1$, and color indicates the angle of the trajectory vector with respect to the canonical variable axes. The μ axes is scaled from -4 to 4, and σ axes are scaled from 0 to 1, where individual dots represent latent encoding values from test images. The rightmost plots show a shift in the relationship between the mean correlation coefficient (MCC) (black, higher is better) and training loss (red, lower is better) as one increases κ.

B CHOOSING A LATENT VARIABLE MODEL

Our proposed method for disentanglement can be implemented in conjunction with different probabilistic latent variable models. In this section, we compare VAEs and normalizing flows as possible candidates.

Variational Autoencoders (VAEs) (Kingma and Welling, 2013) are a widely used probabilistic latent variable model. Despite their simple structure and empirical success, VAEs can converge to a pathological solution called *posterior collapse* (Lucas et al., 2019; Bowman et al., 2016; He et al., 2019). This solution results in the encoder's variational posterior approximation matching the prior, which is typically chosen to be a multivariate standard normal $q(\mathbf{z}|\mathbf{x}) \approx p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. This disconnects the encoder from the decoder, making them approximately independent, i.e. $p(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x})$. The failure mode is often observed when the decoder architecture is overly expressive, i.e. with autoregressive models, or when the likelihood $p(\mathbf{x})$ is easy to estimate. Approaches that alleviate this problem rely on modifying the ELBO training objective (Bowman et al., 2016; Kingma et al., 2016) or restricting the decoder structure (Dieng et al., 2019; Maaløe et al., 2019). However, these approaches come with various drawbacks, including optimization issues (Lucas et al., 2019).

Another approach to estimate latent variables are normalizing flows which describe a sequence of invertible mappings by iteratively applying the change of variables rule (Dinh et al., 2017b). Unlike VAEs, flow based latent variable models allow for a direct optimization of the likelihood (Dinh et al., 2017b). Most normalizing flow models rely on a fast and reliable calculation of the determinant of the Jacobian of the outputs with respect to the inputs, which constrains the architectural design and limits the capacity of the network (Tabak et al., 2010; Tabak and Turner, 2013; Dinh et al., 2017b). Thus, competitive flows require very deep architectures in practice (Kingma and Dhariwal, 2018). Furthermore, flows are not directly suited for a scenario where the observation space is higher dimensional than the generating latent factors, $\dim(\mathbf{z}) < \dim(\mathbf{x})$, as the computation of the determinant requires a square Jacobian matrix. We tried setting $\dim(\mathbf{z}) = \dim(\mathbf{x}) > \dim(\mathbf{z}^*)$, but observed instability while optimizing the objective defined below.

It is straightforward to derive a flow-based objective based on the assumptions in Eq. (2). We consider a normalizing flow with with K blocks $f(\mathbf{x}) = f_K \circ ... \circ f_1 : \mathbf{x} \mapsto \mathbf{z}$. The coupling blocks can refer to nonlinear mixing similar to Kingma and Dhariwal (2018), or in the linear case ($K = 1$) to an invertible de-mixing matrix. This leads to the following estimation of the likelihood

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t) = \; p(f(\mathbf{x}_{t-1}))\; p(f(\mathbf{x}_t)|f(\mathbf{x}_{t-1}))\; \prod_{k=1}^{K} \left|\det \frac{\partial f_k}{\partial \mathbf{z}_{k-1,t-1}}\right|^{-1} \prod_{k=1}^{K} \left|\det \frac{\partial f_k}{\partial \mathbf{z}_{k-1,t}}\right|^{-1}. \tag{23}$$

Note that $p(f(\mathbf{x}_{t-1}))$ is Gaussian and $p(f(\mathbf{x}_t)|f(\mathbf{x}_{t-1}))$ is a Laplacian, similar to Eq. (2). During optimization we take the $-\log$ of both sides and minimize w.r.t. the parameters of f. We refer to this estimator as *SlowFlow*. Our SlowFlow model is very similar to the flow described in (Pineau et al., 2020), who use a Gaussian transition prior and therefore would have weaker identifiability guarantees. Next, we compare SlowFlow and SlowVAE in the context of disentanglement.

To demonstrate the posterior collapse in VAEs, we generate data points $(\mathbf{x}_t, \mathbf{x}_{t-1})$ according to Eq. (2) with a two dimensional latent space $\dim(\mathbf{z}^*) = 2$. We consider a trivial linear mixing of $\mathbf{x}^* = \mathbf{W}^*\mathbf{z}^* = g^*(\mathbf{z}^*)$ with

$$\mathbf{W}^* = \text{diag}(1, \kappa) \tag{24}$$

and $\kappa \in [0.1, 1]$. As can be seen by looking at the σ and μ outputs of the encoder in Fig 6a, for $\kappa < 0.4$, the encoder for the minor axis collapses to the prior. The decoder then tries to minimize the reconstruction loss by solely covering the first principal component of the data, which is also described in Rolinek et al. (2019). Despite the collapse and decrease in MCC, the SlowVAE loss from Eq. (4) still improves during training. On the other hand, a simple linear SlowFlow model $f(\mathbf{x}) = \mathbf{W}\mathbf{x}$, which directly optimizes the likelihood, recovers the latents consistently as seen by the MCC measure (Fig 6b).

To show the strength of the VAE model we increase the complexity of the data-distribution by using a non-linear expanding decoder such that $\dim(\mathbf{x}) \gg \dim(\mathbf{z}^*)$. In Fig. 7 we observe that increasing the input dimensionality is sufficient for SlowVAE to find the corresponding latents and achieve high MCC with low loss.

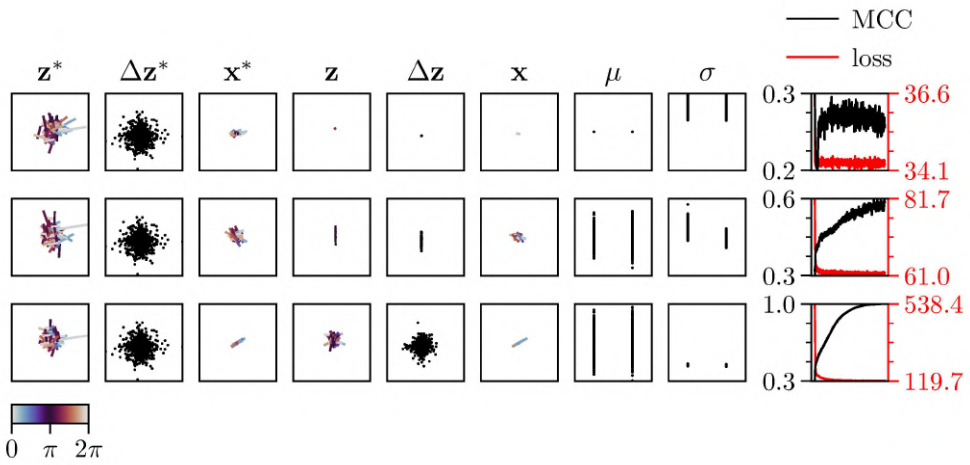


Figure 7: **VAEs perform better when data dimensionality exceeds the latent dimensionality.** VAEs prefer data dimensions to be greater than latent dimensions. Individual subplots are as described in Fig. 6. For all data in this experiment we used a 20-dimensional latent space, $\dim(\mathbf{z}^*) = 20$. Each row corresponds to the dimensionality of the $\mathbf{x}^*$, with values of 20, 200, and 2000. The first two dimensions of $\mathbf{z}^*$ are plotted as well as the two dimensions of $\mathbf{z}$ with the highest corresponding mean correlation coefficient (MCC). The $\mathbf{x}^*$ and $\mathbf{x}$ data are projected onto their first two principal component axes before plotting. A two-layer mixing matrix was used to transform data from Z_{gt} to X_{gt}. As one increases the data dimensionality, the SlowVAE network performs increasingly better in terms of MCC, although worse in terms of total training loss.

Each estimation method is practically useful in different experimental settings. In the case when the mixing operation is trivially defined (Eq. (24), or when the number of dimensions in $\mathbf{z}^*$ match those in $\mathbf{x}^*$), the VAE estimator tends to learn a pathological solution. On the other hand, the normalizing flow estimator does not scale well to high dimensional data due to the requirement of computing the network Jacobian. Additionally, the framework for constructing normalizing flow estimators assumes the latent dimensionality is equal to the data dimensionality to allow for an invertible transform. Together these results lead us to choose an estimator based on the nature of the problem. For our contributed datasets and the DisLib experiments we adopt the VAE framework. However, if one aims to perform simplified experiments such as those typically conducted in the nonlinear ICA literature, it will often make practical sense to switch to a flow-based estimator.

C Disentanglement Metrics

Several recent studies have brought to light shortcomings in a number of proposed disentanglement metrics (Kim and Mnih, 2018; Eastwood and Williams, 2018; Chen et al., 2018; Higgins et al., 2018; Mathieu et al., 2019), many of which have been compiled in the DisLib benchmark. In addition to the concerns they raise, it is important to note that none of the supervised metrics implemented in DisLib allow for continuous ground-truth factors, which is necessary for evaluating with the Natural Sprites and KITTI Masks datasets, as factors such as position and scale are effectively continuous in reality. To rectify this issue without introducing novel metrics, we include the Mean Correlation Coefficient (MCC) in our evaluations, using the implementation of Hyvärinen and Morioka (2016), which is described below.

We measure all metrics presented below between $10,000$ samples of latent factors $\mathbf{z}$ and the corresponding encoded means of our model $\mu(g^*(\mathbf{z}))$. We increase this sample size to $100,000$ for Modularity and MIG to stabilize the entropy estimates.

C.1 MEAN CORRELATION COEFFICIENT

In addition to the DisLib metrics, we also compute the Mean Correlation Coefficient (MCC) in order to perform quantitative evaluation with continuous variables. Because of Theorem 1, perfect disentanglement in the noiseless case should always lead to a correlation coefficient of 1 or -1, although note that we report 100 times the absolute value of the correlation coefficient. In our experiments, MCC is used without modification from the authors' open-sourced code (Morioka, 2018). The method first measures correlation between the ground-truth factors and the encoded latent variables. The initial correlation matrix is then used to match each latent unit with a preferred ground-truth factor. This is an assignment problem that can be solved in polynomial time via the Munkres algorithm, as described in the code release from Morioka (2018). After solving the assignment problem, the correlation coefficients are computed again for the vector of ground-truth factors and the resulting permuted vector of latent encodings, where the output is a matrix of correlation coefficients with D columns for each ground-truth factor and D' rows for each latent variable. We use the (absolute value of the) Spearman coefficient as our correlation measure which assumes a monotonic relationship between the ground-truth factors and latent encodings but tolerates deviations from a strictly linear correspondence.

In the existing implementation for MCC, the ground truth factors, latent encodings, and mixed signal inputs are assumed to have the same dimensionality, i.e. $D = D' = N$. However, in our case, the ground-truth generating factors are much lower dimensional than the signal, $N \ll D$, and the latent encoding is higher dimensional than the ground-truth factors $D' > D$ (see Appendix E for details). To resolve this discrepancy, we add $D' - D$ standard Gaussian noise channels to the ground-truth factors. To compute the MCC score, we take the mean of the absolute value of the upper diagonal of the correlation matrix. The upper diagonal is the diagonal of the square matrix of D ground-truth factors by the top D most correlated latent dimensions after sorting. In this way, we obtain an MCC estimate which averages only over the D correlation coefficients of the D ground truth factors with their corresponding best matching latent factors.

C.2 DISLIB METRICS

BetaVAE (Higgins et al., 2017)

The BetaVAE metric uses a biased estimator with tunable hyperparameters, although we follow the convention established in (Locatello et al., 2018) of using the *scikit-learn* defaults. For a sample in a batch, a pair of images, $(\mathbf{x}_1, \mathbf{x}_2)$, is generated by fixing the value of one of the data generative factors while uniformly sampling the rest. The absolute value of the difference between the latent codes produced from the image pairs is then taken, $\mathbf{z}_{\text{diff}} = |\mathbf{z}_1 - \mathbf{z}_2|$. A logistic classifier is fit with batches of $\mathbf{z}_{\text{diff}}$ variables and the corresponding index of the fixed ground-truth factor serves as the label. Once the classifier is trained, the metric itself is the mean classifier accuracy on a batch of held-out test data. The training minimizes the following loss:

$$L = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \log(\exp(-\mathbf{y}_i(\mathbf{z}_{\text{diff},i}^T\mathbf{w} + c)) + 1), \tag{25}$$

where $\mathbf{w}$ and c are the learnable weight matrix and bias, respectively, and $\mathbf{y}$ is the index of the fixed ground-truth factor for the batch. The network is trained using the `lbfgs` optimizer (Byrd et al., 1995), which is implemented via the *scikit-learn* Python package (Pedregosa et al., 2011) in the Disentanglement Library (DisLib, Locatello et al., 2018). In the original work, the authors argue that their metric improves over a correlation metric such as the mean correlation coefficient by additionally measuring interpretability. However, the linear operation of $\mathbf{z}_{\text{diff},i}^T\mathbf{w} + c$ can perform demixing, which means the measure gives no direct indication of identifiability and thus does not guarantee that the latent encodings are interpretable, especially in the case of dependent factors. Additionally, as noted by Kim and Mnih (2018), BetaVAE can report perfect accuracy when all but one of the ground-truth factors are disentangled, since the classifier can trivially attribute the remaining factor to the remaining latents.

FactorVAE (Kim and Mnih, 2018)

For the FactorVAE metric, the variance of the latent encodings is computed for a large (10,000 in DisLib) batch of data where all factors could possibly be changing. Latent dimensions with variance

below some threshold (0.05 in DisLib) are rejected and not considered further. Next, the encoding variance is computed again on a smaller batch (64 in DisLib) of data where one factor is fixed during sampling. The quotient of these two quantities (with the larger batch variance as the denominator) is then taken to obtain a normalized variance estimate per latent factor. Finally, a majority-vote classifier is trained to predict the index of the ground-truth factor with the latent unit that has the lowest normalized variance. The FactorVAE score is the classification accuracy for a batch of held-out data.

Mutual Information Gap (Chen et al., 2018)

The Mutual Information Gap (MIG) metric was introduced as an alternative to the classifier-based metrics. It provides a normalized measure of the mean difference in mutual information between each ground truth factor and the two latent codes that have the highest mutual information with the given ground truth factor. As it is implemented in DisLib, MIG measures entropy by discretizing the model's latent code using a histogram with 20 bins equally spaced between the representation minimum and maximum. It then computes the discrete mutual information between the ground-truth values and the discretized latents using the *scikit-learn* `metrics.mutual_info_score` function (Pedregosa et al., 2011). For the normalization it divides this difference by the entropy of the discretized ground truth factors.

Modularity (Ridgeway and Mozer, 2018)

Ridgeway and Mozer (2018) measure disentanglement in terms of three factors: modularity, compactness, and explicitness. For modularity, they first measure the mutual information between the discretized latents and ground-truth factors using the same histogram procedure that was used for the MIG, resulting in a matrix, $M \in \mathbb{R}^{D' \times D}$ with entries for each mutual information pair. Their measure of modularity is then

$$\text{modularity} = \frac{1}{D'} \sum_{i=1}^{D'} \Theta \left(1 - \frac{\sum_{j=1}^{D} M_{i,j}^2 - \max(M_i^2)}{\max(M_i^2)(D-1)} \right), \tag{26}$$

where $\max(M_i^2)$ returns the maximum of the vector of squared mutual information measurements between ground truth i and each latent factor. Additionally, Θ is a selection function that returns zero for any i where $\max(M_i^2) = 0$ and otherwise acts as the identity function.

DCI Disentanglement (Eastwood and Williams, 2018)

The DCI scores measure disentanglement, completeness, and informativeness, which have intuitive correspondence to the modularity, compactness, and explicitness of (Ridgeway and Mozer, 2018), respectively. To measure DCI Disentanglement, D regressors are trained to predict each ground truth factor state given the latent encoding. The DisLib implementation uses the `ensemble.GradientBoostingClassifier` function from *scikit-learn* with default parameters, which trains D gradient boosted logistic regression tree classifiers. Importance is assigned to each latent factor using the built-in `feature_importance_` property of the classifier, which computes the normalized total reduction of the classifier criterion loss contributed by each latent. Disentanglement is then measured as

$$\sum_{i=1} D(1 - H(I_i))\tilde{I}_i, \tag{27}$$

where H is the entropy computed with the `stats.entropy` function from *scikit-learn*, $I \in \mathbb{R}^{D \times D'}$ is a matrix of the absolute value of the feature importance between each factor and each ground truth, and $\tilde{I}$ is a normalized version of the matrix

$$\tilde{I}_i = \frac{\sum_{j=1}^{D'} I_{i,j}}{\sum_{k=1}^{D} \sum_{j=1}^{D'} I_{k,j}} \tag{28}$$

SAP Score (Kumar et al., 2018)

To compute the SAP score, Kumar et al. (2018) first train a linear support vector classifier with squared hinge loss and L_2 penalty to predict each ground truth factor from each latent variable. In DisLib this is implemented with the `svm.LinearSVC` function with default parameters from *scikit-learn*. They construct a score matrix $S \in \mathbb{R}^{D' \times D}$, where each entry in the matrix is the

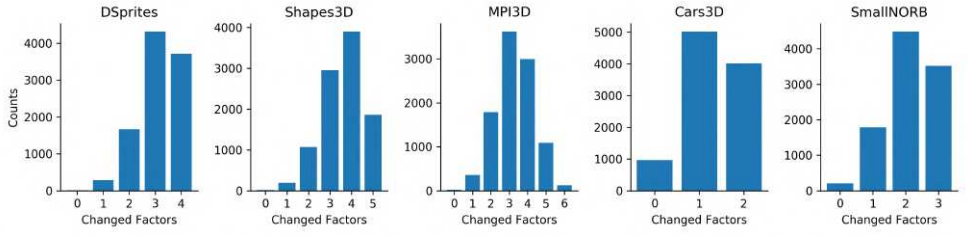


Figure 8: **Number of changing factors in LAP dataset**. For each dataset we sample 10,000 transitions and record the number of changing factors. These are indicated in the histograms. $\lambda = 1$, see Appendix D.

batch-mean classifier accuracy for predicting each ground truth given each individual latent encoding. For each generative factor, they compute the difference between the top two most predictive latent dimensions, which are the two highest scores in a given column of S. The mean (across ground-truth factors) of these differences is the SAP score.

D NATURAL DATASETS

We introduce several datasets to investigate disentanglement in more natural scenarios. Here, we provide an overview on the motivation and design of each dataset.

We have chosen to work with pairs of inputs as minimal sequences because we are interested in the first temporal derivative, more specifically in the sparsity of the transitions between pairs of images. Other methods that look at the second temporal derivative, such as work from Hénaff et al. (2019) on straightening, would require triplets as minimal sequences. Extending our approach beyond this minimal requirement would be simple in terms of the resulting ELBO (which would still factorise like in Eq. 4 because of the Markov property). The only additional complexity would be in the data and loss handling.

An issue with evaluating disentanglement on natural datasets is the fact that the existing disentanglement metrics require knowledge of the underlying generative process of the given data. Although we can observe that the world is composed of distinct entities that vary according to rules imposed by physics, we are unable to determine the appropriate "factors" that generate such scenes. To mitigate this problem, we compile object measurements by calculating the x and y coordinates of the center of mass as well as the area of object masks in natural video frames. We use these measurements to a) inform new disentanglement benchmarks with natural transitions that have similar complexity to existing benchmarks (Natural Sprites) and b) evaluate the ability of algorithms to decode intrinsic object properties (KITTI Masks). We additionally propose a simple extension to the existing DisLib datasets in the form of collecting images into pairs that exhibit sparse (i.e. Laplace) transition probabilities.

D.1 UNIFORM TRANSITIONS (UNI)

The UNI extension is based on the description given by Locatello et al. (2020), where the number of changing factors is determined using draws from a uniform distribution. The key differences between our implementation and theirs is: (i) their code[3] randomly (with 50% probability) sets $k = 1$ even in the $k = \text{Rnd}$ setting, and (ii) we ensure that exactly k factors change. Though we consider these discrepancies minor, we nonetheless label all results reported directly from Locatello et al. (2020) with "LOC", as opposed to "UNI", for clarity.

D.2 LAPLACE TRANSITIONS (LAP)

For each of the datasets in DisLib, we collect pairs of images. For each ground-truth factor, the first value in the pair is chosen from a uniform distribution across all possible values in latent space, while the second is chosen by weighting nearby values in latent space using Laplace distributed probabilities (see Eq. 2). We reject samples that would push a factor outside of the preset range provided by the dataset. We call this the *LAP* DisLib extension. Although the sparse prior indicates that any individual factor is more likely to remain constant, the number of factors that change in a given transition is still typically greater than one. To show this in Fig. 8, we sampled 10,000 transitions from each DisLib dataset with LAP transitions and computed the number of factors that had changed within a pair. This extension of the DisLib datasets provides a bridge from i.i.d. data to natural data by explicitly modeling the observed sparse marginal transition distributions. When training models on the LAP dataset it is possible to reject samples without transitions (i.e. all factors remain constant) since the pair would not result in any temporal learning signal. However, it would arguably be more natural to leave these samples as they would more accurately reflect occurrences of stationary objects in real data. We report the rejection setting in the main text, but found no significant difference between the two settings (see Appendix G).

This dataset also introduces a hyper-parameter λ that controls the rate of the Laplace sampling distribution, while the location is set by the initial factor value. Effectively, when this rate is $\lambda = 1$ most of the factors change most of the time, whereas for a rate of $\lambda = 10$ most of the factors will not change most of the time. Note that this means λ (inversely) changes the scale, which results in larger or smaller movements, but does not affect the distribution itself. In other words, the sparsity is unchanged, as the sparsity is controlled by the shape α. We fix $\lambda = 1$, which yields multiple changes, thus making this dataset fundamentally different both in spirit and in practice, from the UNI dataset.

D.3 YOUTUBE-VOS

For the YouTube dataset, we download annotations from the 2019 version of the video instance segmentation (Youtube-VIS) dataset (Yang et al., 2019)[4], which is built on top of the video object segmentation (Youtube-VOS) dataset (Xu et al., 2018). The dataset has multi-object annotations for every five frames in a 30fps video, which results in a 6fps sampling rate. The authors state that the temporal correlation between five consecutive frames is sufficiently strong that annotations can be omitted for intermediate frames to reduce the annotation efforts. Such a skip-frame annotation strategy enables scaling up the number of videos and objects annotated under the same budget, yielding 131,000 annotations for 2,883 videos, with 4,883 unique video object instances. Although we do not evaluate against YouTube-VOS in this study, we see it as the logical next step in transitioning to natural data. The large scale, lack of environmental constraints, and abundance of object types makes it the most challenging of the datasets considered herein.

The original image size of the YouTube-VOS dataset is 720×1280. In order to preserve the statistics of the transitions, we choose not to directly downsample to 64×64, but instead preserve the aspect ratio by downsampling to 64×128. In order to minimize the bias yielded by the extraction method, noting the center bias typically present in human videos, we extract three overlapping, equally spaced 64×64 pixel windows with a stride of 32. For each resulting $64 \times 64 \times T$ sequence, where T denotes the number of time steps in the sequence, we filter out all pairs where the given object instance is not present in adjacent frames, resulting in 234,652 pairs.

D.4 NATURAL SPRITES

The benchmark is available at `https://zenodo.org/record/3948069`.

Without a metric for disentanglement that can be applied to unknown data generating processes, we are limited to synthetic datasets with known ground-truth factors. Let us take dSprites (Matthey et al., 2017) as an example. The dataset consists of all combinations of a set of latent factor values, namely,

- Color: white

[3] `https://github.com/google-research/disentanglement_lib/blob/master/disentanglement_lib/methods/weak/train_weak_lib.py#L48`

[4] `https://competitions.codalab.org/competitions/20127`

Config	Scale	X	Y	(R, G, B)	Shape	Orientation
Continuous	YT [2375]	YT [197342]	YT [187112]	(1.0, 1.0, 1.0)	(square, triangle, star_4, spoke_4)	(0,9,...,342,351)
Discrete	YT [6]	YT [32]	YT [32]	(1.0, 1.0, 1.0)	(square, triangle, star_4, spoke_4)	(0,9,...,342,351)

Table 4: Natural Sprite Configs. Values in brackets refer to the number of unique values. Shapes presented are predefined in Spriteworld (Watters et al., 2019).

- Shape: square, ellipse, heart
- Scale: 6 values linearly spaced in $[0.5, 1]$
- Orientation: 40 values in $[0, 2\pi]$
- Position X: 32 values in $[0, 1]$
- Position Y: 32 values in $[0, 1]$

Given the limited set of discrete values each factor can take on, all possible samples can be described by a tractable dataset, compiled and released to the public. But, in reality, all of these factors should be continuous: a spectrum of possible colors, shapes, scales, orientations, and positions exist. We address this by constructing a dataset that is augmented with natural and continuous ground truth factors, using the mask properties measured from the YouTube dataset described in Appendix D.3.

We can choose the complexity of the dataset by discretizing the 234,652 transition pairs of position and scale into an arbitrary number of bins. In this study, we discretize to match the number of possible object states as dSprites, which we present in Table 4. This helps isolate the effect of including natural transitions from the effect of increasing data complexity. We produce a pair by fixing the color, shape, and orientation, but updating the position and scale with transitions sampled from the YouTube measurements. We motivate fixing shape and color by noting that this is consistent with object permanence in the real world. We decided to fix the orientation because we do not currently have a way to approximate it from object masks and we did not want to introduce artificial transition probabilities. To minimize the effect of extreme outliers, we filter out 10% of the data by removing frames if the mask area falls below the 5% or above the 95% quantiles, which reduces the number of pairs to 207,794. Finally, we use the Spriteworld (Watters et al., 2019) renderer to generate the images. Spriteworld allows us to render entirely new sprite objects at the precise position and scale as was measured from YouTube. For example, if one would want to apply YouTube-VOS transitions to MPI3D (Gondal et al., 2019), this option is unavailable without the associated renderer.

In relation to the Laplace transitions described in section D.2, this update i) produces pairs that correspond to transitions observed in real data, ii) allows for smooth transitions by defining the data generation process as opposed to being limited by the given collected dataset (e.g. dSprites), and iii) includes complex dependencies among factors that are present in natural data. We generate the data online, thus training the model to fit the underlying distribution as opposed to a sampled finite dataset.

However, as noted previously, all supervised metrics aggregated in DisLib are inapplicable to continuous factors, which is problematic as the generating distribution is effectively continuous with respect to a subset of the factors. Therefore, we limit our quantitative evaluation to MCC for continuous datasets. However, we are able to evaluate disentanglement with the standard metrics on the discretized version.

D.5 KITTI MOTS Pedestrian Masks (KITTI Masks)

The benchmark is available at `https://zenodo.org/record/3931823`.

While Natural Sprites enables evaluation of disentanglement with natural transitions, we note that any disentanglement framework that requires knowledge of the underlying generative factors is unrealistic for real-world data. Measurements such as scale and position correspond to object properties that are ecologically relevant to the observer and can serve as suitable alternatives to the typical generative factors. We directly test this using our KITTI Masks dataset.

To create the dataset, we download annotations from the Multi-Object Tracking and Segmentation (MOTS) Evaluation Benchmark (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016),

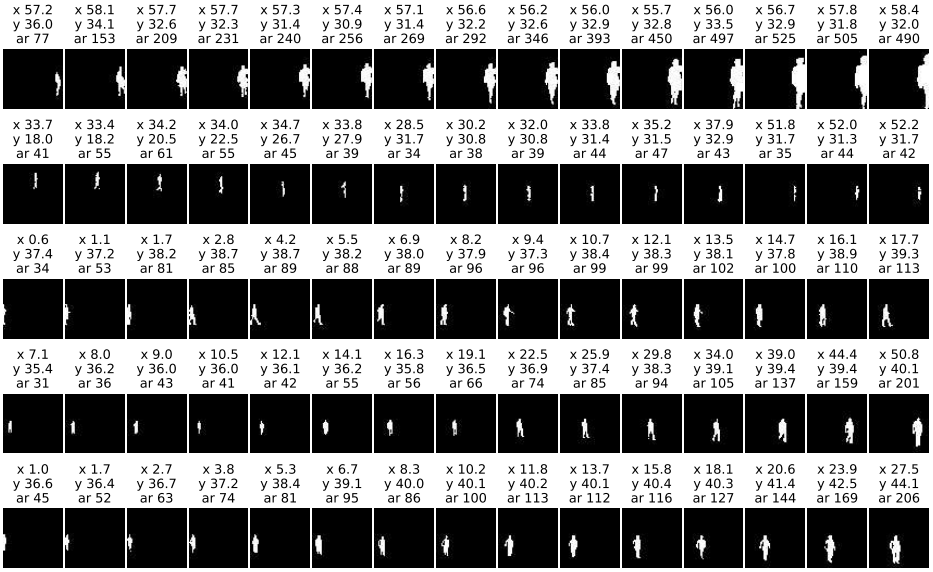


Figure 9: **KITTI Masks**. Each row corresponds to sequential frames from random sequences in the KITTI Mssks dataset. Above each image we denote measured object properties where x, y correspond the center of mass position and ar corresponds to the area.

which is split into KITTI MOTS and MOTSChallenge[5]. Both datasets contain sequences of pedestrians with their positions densely annotated in the time and pixel domains. For simplicity, we only consider the instance segmentation masks for pedestrians and do not use the raw data.

The resulting KITTI Masks dataset consists of 2,120 sequences of individual pedestrians with lengths between 2 and 710 frames each, resulting in a total of 84,626 individual frames. As we did with YouTube-VOS, we estimate ground truth factors by calculating the x and y coordinates of the center of mass of each pedestrian mask in each frame. We define the object size as the area of the mask, i.e. the total number of pixels. We consider the disentanglement performance for different mean time gaps between image pairs in table 2 and Appendix G.3. For samples and the corresponding ground truth factors see Fig. 9.

The original KITTI image sizes are 1080×1920 or 480×640 resolution for MOTSChallenge and between 370 and 374 pixels tall by 1224 and 1242 pixels wide for KITTI MOTS. The frame rates of the videos vary from 14 to 30 fps, which can be seen in Table 2 of Milan et al. (2016). We use nearest neighbor down-sampling for each frame such that the height was 64 pixels and the width is set to conserve the aspect ratio. After down-sampling, we use a horizontal sliding window approach to extract six equally spaced windows of size 64×64 (with overlap) for each sequence in both datasets. This results in a $64 \times 64 \times T$ sequence, where T denotes the number of time steps in the sequence. Note that here we make reasonable assumptions on horizontal translation and scale invariance of the dataset. We justify the assumed scale invariance by observing that the data is collected from a camera mounted onto a car which has varying distance to pedestrians. To confirm the translation invariance, we performed an ablation study on the number of horizontal images. Instead of six horizontal, equally spaced sliding windows, we only use two which leads to differently placed windows. We do not observe significant changes in the reported data statistics (e.g. the kurtosis of the fit stays within $\pm 10\%$ of the previous value for Δx transitions). The values of Δy and $\Delta area$ do not change significantly compared to Table 7.

For each resulting $64 \times 64 \times T$ sequence, where T denotes the number of time steps in the sequence, we extract all individual pedestrian masks based on their object instance identity and create a new sequence for each pedestrian such that each resulting sequence only contains a single pedestrian. We ignore images with masks that have less than 30 pixels as they are too far away or occluded and were

[5] https://www.vision.rwth-aachen.de/page/mots

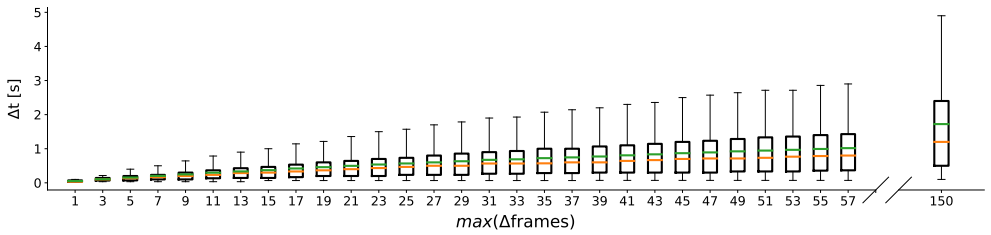


Figure 10: **KITTI Masks Δt**. Boxes indicate correspondence to physical time for different max(Δframes) in the KITTI Masks datasets. The orange line denotes the median and the green line the mean. The whiskers cover the 5th and 95th percentile of data.

not recognizable by the authors. We keep all sequences of two or more frames, as the algorithm only requires pairs of frames for training.

We leave the maximum distance between time frames within a pair, max(Δframes), as a hyperparameter. For a given max(Δframes), we report the mean change in physical time in seconds (denoted by mean(Δt)). We test adjacent frames (max(Δframes) $= 1$), which corresponds to a mean($\Delta t = 0.05$) and max(Δframes) $= 5$, which corresponds to a mean($\Delta t = 0.15$). This procedure is motivated by the fact that different sequences were recorded with different frame rates and reporting the mean(Δt) in seconds allows for a physical interpretation. The relationship between max(Δframes) and mean(Δt) is in Fig. 10. We show results for testing additional values of mean(Δt) in Appendix G.3.

During training, we augment the data by applying horizontal and vertical translations of ± 5 pixels and rotations of $\pm 2^\circ$ degree. We apply the exact same data augmentation to both images within a pair to not change any transition statistics.

We note that both YouTube-VOS (Xu et al., 2018; Yang et al., 2019) and KITTI-MOTS (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016) are multi-object datasets, although we consider each unique object (mask) separately. Multi-object representation learning and disentanglement are highly connected, in fact they have recently begun to be used interchangeably (Wulfmeier et al., 2020).

To briefly comment on possible extensions in this direction, we see no reason why our prior would not be beneficial to multi-object methods such as MONet (Burgess et al., 2019) and IODINE (Greff et al., 2019), or video extensions such as ViMON (Weis et al., 2020) and OP3 (Veerapaneni et al., 2019).

E Model Training and Selection

We train all models on all datasets provided in DisLib with the UNI and LAP variants.

All models are implemented in PyTorch (Paszke et al., 2019). To facilitate comparison, the training parameters, e.g. optimizer, batch size, number of training steps, as well as the VAE encoder and decoder architecture are identical to those reported in (Locatello et al., 2018; 2020). We use this architecture for all datasets, only adjusting the number of input channels (greyscale for dSprites, smallNORB, and KITTI Masks; three color channels for all other datasets).

The model formulation is agnostic to the direction of time. Therefore, to increase the temporal training signal at a fixed computational cost for each batch of input pairs $(\mathbf{x}_0, \mathbf{x}_1)$, we optimize the model in both directions i.e. optimizing the model objective for both $t_0 = 0$, $t_1 = 1$ as well as $t_0 = 1$, $t_1 = 0$.

F EXTENDED COMPARISONS AND CONTROLS

F.1 COMPARISON TO NONLINEAR ICA

F.1.1 THEORETICAL COMPARISON

Nonlinear ICA has recently been advanced significantly by several papers from Hyvärinen and colleagues. Of these studies, the two that are most comparable to our work is Hyvärinen and Morioka (2017), which uses an unsupervised contrastive loss for nonlinear demixing and Khemakhem et al. (2020a), which extends the nonlinear ICA framework to include variational autoencoders (VAEs). However, our theory covers an important class of transitions relevant for natural data that is not covered by the identifiability proofs of either of the aforementioned studies.

As a specific comparison to the first paper, the non-Gaussian autoregressive model that their identifiability proof rests upon (Eq. 8 in Hyvärinen and Morioka, 2017) assumes that the second derivative of the innovation probability density function is less than zero to satisfy *uniform dependence*, which is only met for $\alpha > 1$ for generalized Laplace transition distributions. While they denote (footnote 3) that Laplace distributions ($\alpha = 1$) are not covered by their theory, they offer a suggestion for a smooth approximation. However, they do not demonstrate that this approximation is useful in practice, or offer a solution to a general class of sparse distributions for $\alpha \leq 1$. We chose a generalized Laplacian to fit our data and for our model assumption as it allows for simple parameterization of fits to data (e.g. $\alpha = 0.5$ for natural movie transitions), but is simultaneously quite expressive (Sinz et al., 2009). Though we use $\alpha = 1$ in practice for our estimation method, we prove identifiability up to permutations and sign flips for any $\alpha < 2$, covering all sparse distributions under the expressive generalized Laplacian model. In addition, we assume a Gaussian marginal distribution that allows us to derive a fundamentally stronger proof of identifiability – where we identify up to permutation and sign-flips. Hyvärinen and Morioka (2017) only identify the sources up to arbitrary non-linear element-wise transformations. Thus they require a subsequent step of ICA (under the typical assumption that at most one marginal source distribution is Gaussian) to recover the signal up to permutations and sign flips for a class of distributions where it is unclear whether they account for temporal sparsity.

The work of Khemakhem et al. (2020a) has a couple of differences from our own, most notable of which is the form of the conditional prior, $p(\mathbf{z}_t|\mathbf{z}_{t-1})$. They assume that the conditional posterior is part of the exponential family, which does not include Laplacian conditionals. Though the exponential family contains the Laplace distribution with fixed mean as its member, it does not allow their approach to model sparse transitions. They assume that the natural parameters of the exponential family distribution are conditioned on $\mathbf{z}_{t-1}$, meaning that only the scale but not the mean of the Laplace prior for $\mathbf{z}_t$ can be modulated by the previous time step, thus not allowing for sparse transition probabilities. Additionally, their implementation requires the number of classes (i.e. states of the conditioning variable) to equal the number of stationary segments, which is impractical for the datasets we consider.

Thus, we provide a closer match to natural data transitions, with a stronger identifiability result. We provide validation by performing an extensive evaluation leveraging our contributed datasets as well as the models, metrics, and datasets provided by the Disentanglement Library (DisLib, discussed in section 4). We consider methods from the disentanglement literature (Locatello et al., 2020) as well as nonlinear ICA (Hyvärinen and Morioka, 2017), that are functionally capable of processing transitions.

F.1.2 EMPIRICAL COMPARISON

Hyvärinen and Morioka (2017) conducted a simulation where the sources in the nonlinear ICA model come from a linear autoregressive (AR) model with non-Gaussian innovations. Specifically, temporally dependent 20-dimensional source signals were randomly generated according to $\log p(s(t)|s(t-1)) = -|s(t) - 0.7s(t-1)|$. Though this generative process was noted to not be covered by the theory presented in (Hyvärinen and Morioka, 2017), the authors demonstrated that PCL could reconstruct the source signals reasonably well even for the nonlinear mixture case. Given our practical use of a Laplacian conditional, we found it a valuable comparison to evaluate our theory in this artificial setting.

Method	L=1	L=2	L=3	L=4	L=5
PCL	**0.998**	0.960	0.950	0.917	0.902
PCL (NF)	0.946	0.918	0.918	0.917	0.876
SlowFlow	0.997	**0.987**	**0.982**	**0.975**	**0.975**

Table 5: MCC using linear correlation where L denotes the number of mixing layers.

Given the discussion in Appendix B, we use SlowFlow for these experiments. For computational tractability in demixing highly nonlinear transformations, we consider normalizing flows (Dinh et al., 2017a;b; Kingma and Dhariwal, 2018), namely volume-preserving flows (Sorrenson et al., 2017), as we find constraining the Jacobian determinant stabilizes learning. To ensure sufficient expressivity, we consider 6 coupling blocks, each containing a 2-layer MLP with 500 hidden units and ReLU nonlinearities. We compare to the PCL implementation presented in (Hyvärinen and Morioka, 2017), where an MLP with the same number of hidden layers as the mixing MLP was adopted. We use 100 hidden units as we did not find increasing the value improved performance. To account for the architectural difference serving as a possible confounder, we use the same normalizing flow encoder for optimizing the PCL objective, which we term "PCL (NF)".

While (Hyvärinen and Morioka, 2017) used leaky ReLU nonlinearities to make the mixing invertible, said mixing is non-differentiable. This is problematic for SlowFlow, as it involves gradient optimization of the Jacobian term, and more importantly, unlike PCL, aims to explicitly recover the mixing process. We thus use a a smooth version of the leaky-ReLU activation function with a hyperparameter α (Gresele et al., 2020),

$$\mathrm{s_L}(x) = \alpha x + (1-\alpha)\log(1+e^x). \tag{29}$$

By ensuring the mixing process is smooth, we find that SlowFlow performs favorably relative to PCL (Table 5) when evaluated in the same setting, converging to a better optimum at higher levels of mixing.

F.2 Joint Factor Dependence Evaluation

In order to consider joint dependencies among natural generative factors, we leverage Natural Sprites to construct modified datasets where time-pairs of factors are shuffled per-factor (e.g. combining the x transition from one clip with the y transition from a different clip). This destroys dependencies between the factors, while maintaining the sparse marginal distributions. In Fig. 11 (right), we show 2D marginals before (blue) and after (orange) this shuffling. The additional density on the diagonals in the unshuffled data reveals dependencies between pairs of factors on both datasets. As mentioned in section 3.4, the observed dependency is mismatched from the theoretical assumptions of our model.

We test how robust SlowVAE is to such a mismatch by training it on the *permuted* data and re-evaluating disentanglement. In Table 22, we highlight that the improvement of SlowVAE on the permuted (i.e. independent) continuous Natural Sprites is not significant. In Table 21, we surprisingly find an overall improved score with non-permuted transitions (i.e. with dependencies), with three out of seven metrics showing a significant improvement. This is in line with Fig. 1f in Khemakhem et al. (2020b), where, at least for simple mixing, a model (Khemakhem et al., 2020a) that does not account for dependencies performs as well as one that does (Khemakhem et al., 2020b). We conclude that these preliminary results do not support the hypothesis that SlowVAE's disentanglement is reliant upon the model assumption that the factors are independent, but do acknowledge that the empirical effect of statistical dependence in natural video warrants further exploration (Träuble et al., 2020; Yang et al., 2020).

F.3 Transition Prior Ablation

We consider an ablated model which minimizes a KL-divergence term between the posteriors at time-step t and time-step $t-1$. This encourages the model to match the posteriors of both time points as closely as possible, and resembles a probabilistic variant of Slow Feature Analysis (Turner and Sahani, 2007). Specifically, we set $p(\mathbf{z}_t|\mathbf{z}_{t-1}) = q(\mathbf{z}_{t-1}|\mathbf{x}_{t-1})$, replacing the Laplace prior with the

posterior of the previous time step. This is equivalent to a Gaussian ($\alpha = 2$) transition prior, where the mean and variance are specified by the previous time step. We ablate over the regularization parameter γ and provide results in Tables 14 and 15, although we note that we still use the same hyperparameter values for SlowVAE as in all other experiments. As predicted by our theoretical result, $\alpha = 2$ leads to *entangled* representations in aggregate across evaluated datasets and metrics, even when considering a spectrum of γ values, resulting in a drastic reduction in scores, particularly on dSprites and Natural Sprites.

G Additional Results

G.1 Extended Data Analysis

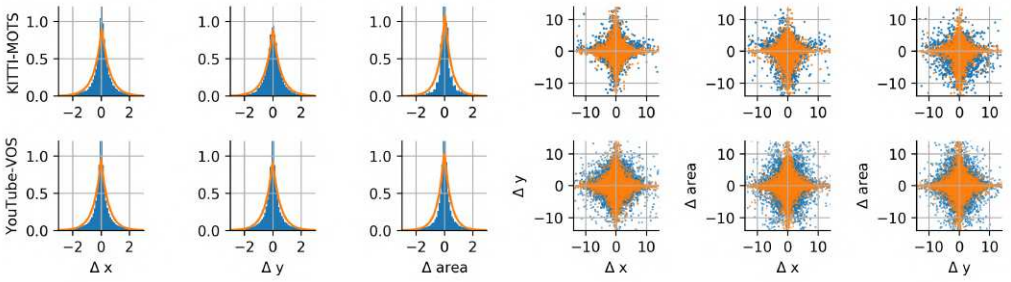


Figure 11: **Statistics of Natural Transitions**. Left) Distribution over transitions for horizontal (Δx) and vertical (Δy) position as well as mask/object size ($\Delta area$) for both datasets. Orange lines indicate fits of generalized Laplace distributions (Eq. 2). Right) 2D marginal distribution over pairs of factor transitions (blue) and permuted pairs (orange) that indicate the marginal distributions when made independent.

dataset	N	Δ area	Δ x	Δ y
KITTI-MOTS	82506	0.45	0.59	0.69
YouTube-VOS	234652	0.44	0.52	0.55

Table 6: Shape parameters (α) of the fitted generalized Laplace distributions in Fig. 11.

We report the empirical estimates of Kurtosis in Table 7. We report the log-likelihood scores for the Δ area, Δ x, Δ y statistics in Tables 8, 9, and 10, respectively for a Normal, a Laplace and a generalized Laplace/Normal distribution. For these distributions, we also report the fit parameters for the Δ area, Δ x, Δ y statistics in Tables 11, 12, and 13, respectively, where the shape parameter α of the generalized Laplacian is in bold face. As a higher likelihood indicates a better fit, we can see further evidence that natural transitions are highly leptokurtic; a Laplace distribution ($\alpha = 1$) is a better fit than a Gaussian ($\alpha = 2$), while the generalized Laplacian yields the highest likelihood consistently with $\alpha \approx 0.5$ for all measurements, as indicated in the main paper. For the plots in Figs. 1 and 11, we set the standard deviation of each component to 1 and clipped the minimum (-5) and maximum (5) values.

We note that while the marginal transitions appear sparse in metrics computed from the given object masks, our analysis considers 2D projections of objects instead of the transition statistics in their 3D environment. Understanding the relationship between 3D and 2D transition statistics is a compelling question from a broader perspective of visual processing, but unfortunately, the KITTI-MOTS masks (Voigtlaender et al., 2019; Geiger et al., 2012; Milan et al., 2016) lack the associated depth data required to answer it. Nonetheless, the natural scene statistics we compute are relevant, given that most computer vision models and vision-based animals see the 3D world as projected onto their 2D receptor arrays.

dataset	N	Δ area	Δ x	Δ y
KITTI	82506	68.92	38.50	65.39
YouTube	234652	76.49	39.98	35.59

Table 7: Empirical estimates of Kurtosis for mask transitions per metric for each dataset.

dataset	N	genlaplace	normal	laplace
KITTI	82506	-3.21e+05	-3.79e+05	-3.35e+05
YouTube	234652	-1.29e+06	-1.45e+06	-1.33e+06

Table 8: Maximum likelihood scores for the considered distributions on Δ **area** for each dataset.

dataset	N	genlaplace	normal	laplace
KITTI	82506	-8.72e+04	-1.20e+05	-9.25e+04
YouTube	234652	-4.50e+05	-5.64e+05	-4.74e+05

Table 9: Maximum likelihood scores for the considered distributions on Δx for each dataset.

dataset	N	genlaplace	normal	laplace
KITTI	82506	-7.59e+04	-1.07e+05	-7.86e+04
YouTube	234652	-4.40e+05	-5.45e+05	-4.60e+05

Table 10: Maximum likelihood scores for the considered distributions on Δy for each dataset.

dataset	N	genlaplace	normal	laplace
KITTI	82506	[**4.55e-01**, 1.00e+00, 1.01e+00]	[4.53e-01, 2.39e+01]	[1.00e+00, 1.07e+01]
YouTube	234652	[**4.44e-01**, 1.47e-16, 5.04e+00]	[2.25e-01, 1.16e+02]	[7.73e-09, 5.28e+01]

Table 11: Parameter fits for the considered distributions on Δ **area** for each dataset. The parameters are (alpha, location, scale) for generalized Laplace/Normal, (location, scale) for the other two distributions.

dataset	N	genlaplace	normal	laplace
KITTI	82506	[**5.87e-01**, 4.76e-02, 1.69e-01]	[5.34e-02, 1.04e+00]	[5.49e-02, 5.64e-01]
YouTube	234652	[**5.15e-01**, 1.15e-14, 2.57e-01]	[2.32e-03, 2.68e+00]	[7.54e-09, 1.38e+00]

Table 12: Parameter fits for the considered distributions on Δ **x** for each dataset. The parameters are (alpha, location, scale) for generalized Laplace/Normal, (location, scale) for the other two distributions.

dataset	N	genlaplace	normal	laplace
KITTI	82506	[**6.94e-01**, 1.02e-02, 2.32e-01]	[3.84e-02, 8.86e-01]	[1.71e-02, 4.77e-01]
YouTube	234652	[**5.48e-01**, 2.93e-13, 3.08e-01]	[8.81e-03, 2.47e+00]	[9.15e-04, 1.30e+00]

Table 13: Parameter fits for the considered distributions on Δ **y** for each dataset. The parameters are (alpha, location, scale) for generalized Laplace/Normal, (location, scale) for the other two distributions.

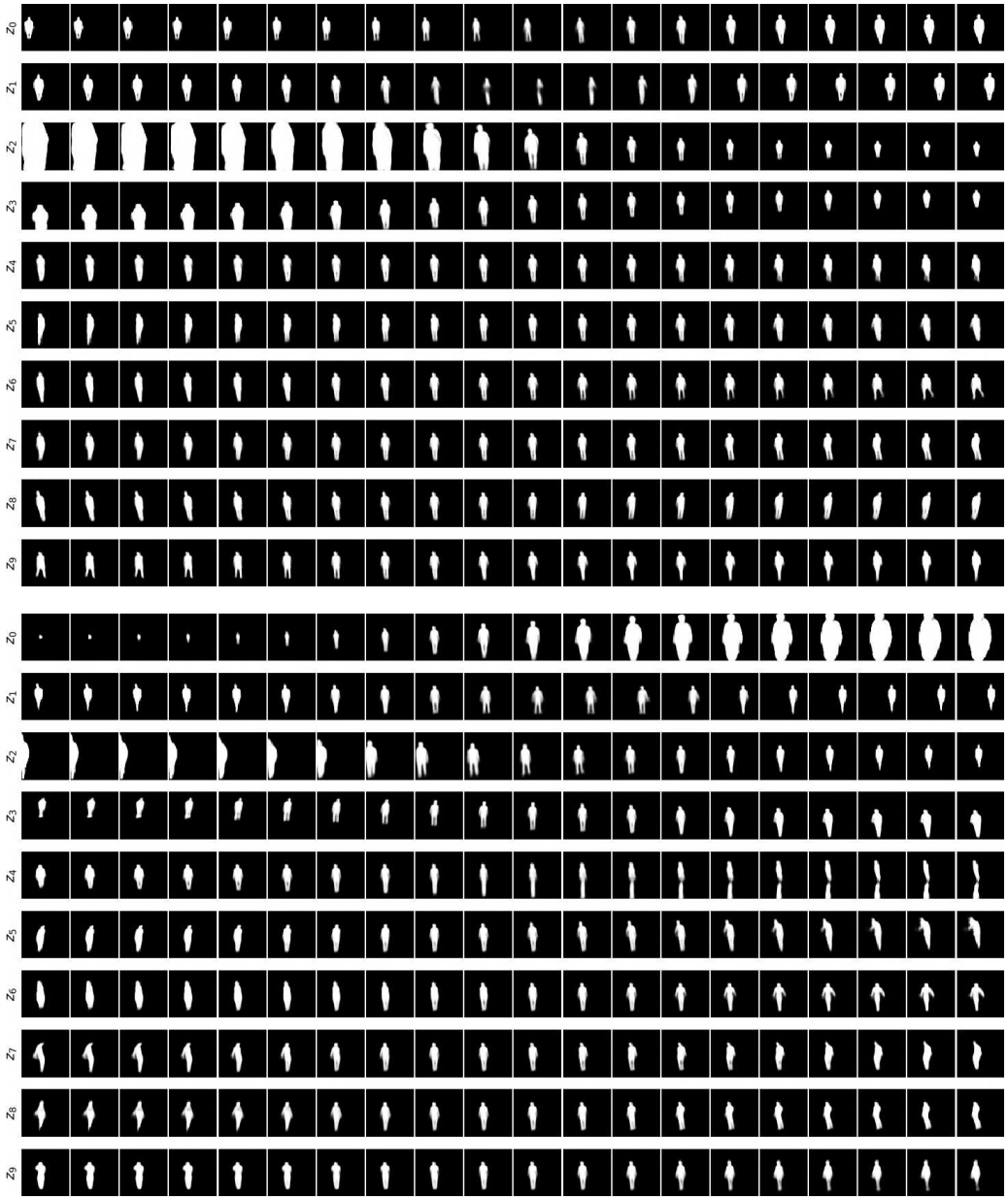


Figure 12: **KITTI Masks Latent Representations.** We show axis latent traversals along each dimension for the β-VAE (top) and SlowVAE (bottom). Here, the latents z_i are sorted from top to bottom in ascending order according to the mean variance output of the encoder. With MCC correlation (see e.g. Fig. 20) the known ground truth factors are matched as following: β-VAE: scale$\sim$ z_2, x-position$\sim$ z_1 and y-position$\sim$ z_3; SlowVAE: scale$\sim$ z_0, x-position$\sim$ z_1 and y-position$\sim$ z_3. With these latent visualizations alone, there is no significant difference visible between β-VAE and SlowVAE. However, we see a quantitative difference with the MCC score (see Table 2) and a qualitative difference when directly observing latent embeddings (see Fig. 20).

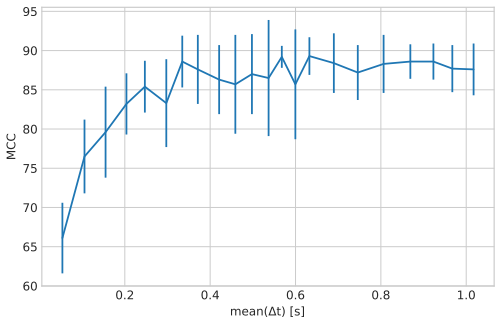


Figure 13: Ablation over mean(Δt) for SlowVAE. Mean and standard deviation (s.d.) MCC scores

Model	Data	BetaVAE	FactorVAE	MIG	MCC	DCI	Modularity	SAP
SlowVAE	dSprites (Laplace)	100.0 (0.0)	97.5 (3.0)	29.5 (9.3)	69.8 (2.3)	65.4 (3.6)	96.5 (1.6)	8.1 (3.0)
PM-VAE (16)	dSprites (Laplace)	64.1 (7.0)	44.8 (13.0)	5.2 (2.3)	45.0 (5.5)	5.9 (3.9)	93.5 (1.9)	1.7 (0.8)
PM-VAE (10)	dSprites (Laplace)	78.8 (7.5)	59.4 (11.2)	5.9 (1.8)	49.2 (4.3)	13.6 (5.6)	92.7 (3.0)	3.9 (1.7)
PM-VAE (8)	dSprites (Laplace)	82.9 (2.8)	61.2 (5.7)	7.1 (2.6)	49.6 (3.3)	14.5 (3.5)	91.6 (3.0)	4.3 (1.6)
PM-VAE (4)	dSprites (Laplace)	86.6 (2.7)	64.1 (7.2)	11.6 (5.0)	52.0 (3.8)	22.9 (3.7)	90.9 (2.7)	5.7 (2.8)
PM-VAE (2)	dSprites (Laplace)	86.3 (2.4)	62.9 (7.7)	10.9 (3.2)	50.0 (3.5)	21.2 (5.3)	92.3 (1.9)	5.5 (2.0)
PM-VAE (1)	dSprites (Laplace)	82.5 (5.4)	58.4 (6.0)	7.6 (3.6)	45.9 (4.9)	14.4 (5.1)	92.1 (4.0)	4.0 (2.0)
SlowVAE	Natural (Discrete)	82.6 (2.2)	76.2 (4.8)	11.7 (5.0)	52.6 (4.1)	18.9 (5.5)	88.1 (3.6)	4.4 (2.3)
PM-VAE (16)	Natural (Discrete)	72.7 (2.8)	49.2 (3.7)	2.8 (1.2)	38.3 (3.2)	6.9 (1.8)	85.3 (1.8)	1.2 (0.7)
PM-VAE (10)	Natural (Discrete)	76.6 (3.6)	52.0 (4.9)	3.8 (2.2)	39.0 (3.9)	7.3 (1.8)	87.0 (2.2)	2.0 (1.0)
PM-VAE (8)	Natural (Discrete)	74.6 (3.4)	49.3 (4.4)	3.1 (1.8)	38.9 (3.2)	7.1 (1.8)	87.8 (1.7)	1.6 (1.0)
PM-VAE (4)	Natural (Discrete)	73.8 (3.8)	48.8 (5.3)	2.7 (1.5)	35.7 (3.5)	6.7 (2.0)	87.4 (2.2)	1.6 (0.9)
PM-VAE (2)	Natural (Discrete)	73.4 (3.1)	47.0 (5.3)	2.2 (1.1)	36.8 (2.4)	6.2 (1.5)	87.4 (1.9)	1.1 (0.6)
PM-VAE (1)	Natural (Discrete)	73.5 (3.3)	49.7 (5.4)	3.1 (1.6)	36.9 (3.2)	6.9 (1.8)	86.9 (2.2)	1.8 (0.7)

Table 14: Mean and standard deviation (s.d.) metric scores across 10 random seeds. PM-VAE (γ) refers to replacing the Laplace prior with a KL-divergence term between the (Gaussian) posteriors at time-step t and time-step $t-1$, with conditional prior regularization, γ.

G.2 All DisLib Results

We include results on all DisLib datasets, dSprites (Matthey et al., 2017), Cars3D (Reed et al., 2015), SmallNORB (LeCun et al., 2004), Shapes3D (Kim and Mnih, 2018), MPI3D (Gondal et al., 2019), in Tables 16, 17, 18, 19, and 20, respectively. We report both median (a.d.) to compare to the previous median scores reported in (Locatello et al., 2020), as well as the the more common mean (s.d.) scores for future comparisons and straightforward statistical estimates of significant differences between models. We also consider allowing for static transitions, which we denote with "NC", e.g. "LAP-NC", in the tabular results. As mentioned in Section 5, we use the same parameter settings for SlowVAE in all experiments, while model selection was performed not only per dataset, but per seed, for results from (Locatello et al., 2020).

G.3 KITTI Masks Δt Ablation

As seen in the main text, considering image pairs separated further apart in time appears beneficial. Here we evaluate a wider range by taking frames which are further apart in a sequence. max(Δframes) $= N$ indicates that all pairs differ by *at most* N frames. We chose an upper bound of N, rather than sampling pairs with a fixed separation, to account for the variable frame rates and sequence lengths in the original dataset (Milan et al., 2016) without introducing a confounding factor of varying dataset size. We report in Fig. 10 how the max(Δframes) criterion corresponds to the mean time gap between image pairs (mean(Δt)) in seconds. For further details, we refer to Appendix D.5.

In Fig. 13 we visualize an ablation over mean(Δt). We find that model performance increased initially with larger temporal separation between data points, then plateaued. We also observe in Fig. 14 that the measured factor marginals remain sparse, with $\alpha < 1$, for all tested settings of mean(Δt).

Model	Data	MCC
SlowVAE	Natural (Continuous)	49.1 (4.0)
PM-VAE (16)	Natural (Continuous)	35.2 (3.7)
PM-VAE (10)	Natural (Continuous)	33.2 (2.1)
PM-VAE (8)	Natural (Continuous)	32.7 (3.1)
PM-VAE (4)	Natural (Continuous)	33.7 (2.3)
PM-VAE (2)	Natural (Continuous)	32.4 (3.2)
PM-VAE (1)	Natural (Continuous)	34.2 (3.4)
SlowVAE	Kitti (mean$(\Delta t) = 0.05s$)	66.1 (4.5)
PM-VAE (16)	Kitti (mean$(\Delta t) = 0.05s$)	63.1 (9.3)
PM-VAE (10)	Kitti (mean$(\Delta t) = 0.05s$)	57.4 (8.5)
PM-VAE (8)	Kitti (mean$(\Delta t) = 0.05s$)	59.0 (5.6)
PM-VAE (4)	Kitti (mean$(\Delta t) = 0.05s$)	51.8 (9.2)
PM-VAE (2)	Kitti (mean$(\Delta t) = 0.05s$)	50.3 (7.4)
PM-VAE (1)	Kitti (mean$(\Delta t) = 0.05s$)	38.4 (6.8)
SlowVAE	Kitti (mean$(\Delta t) = 0.15s$)	79.6 (5.8)
PM-VAE (16)	Kitti (mean$(\Delta t) = 0.15s$)	69.6 (5.9)
PM-VAE (10)	Kitti (mean$(\Delta t) = 0.15s$)	78.2 (6.0)
PM-VAE (8)	Kitti (mean$(\Delta t) = 0.15s$)	73.8 (10.0)
PM-VAE (4)	Kitti (mean$(\Delta t) = 0.15s$)	67.9 (10.4)
PM-VAE (2)	Kitti (mean$(\Delta t) = 0.15s$)	60.7 (8.8)
PM-VAE (1)	Kitti (mean$(\Delta t) = 0.15s$)	60.9 (9.1)

Table 15: Continuous ground-truth variable datasets. See Table 14 for details.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	82.3	66.0	10.2	18.6	82.2	4.9
Ada-ML-VAE (LOC)	89.6	70.1	11.5	29.4	89.7	3.6
Ada-GVAE (LOC)	92.3	84.7	26.6	47.9	91.3	7.4
SlowVAE (UNI)	89.7 (3.8)	81.4 (8.4)	34.5 (9.6)	50.0 (6.9)	87.1 (2.0)	5.1 (1.5)
SlowVAE (LAP)	100.0 (0.0)	99.2 (2.3)	28.2 (8.2)	65.5 (3.1)	96.8 (1.4)	6.0 (2.4)
SlowVAE (LAP-NC)	100.0 (0.2)	97.4 (4.4)	29.1 (7.1)	62.0 (4.2)	97.4 (1.6)	8.2 (2.9)
SlowVAE (UNI)	87.0 (5.1)	75.2 (11.1)	28.3 (11.5)	47.7 (8.5)	86.9 (2.8)	4.4 (2.0)
SlowVAE (LAP)	100.0 (0.0)	97.5 (3.0)	29.5 (9.3)	65.4 (3.6)	96.5 (1.6)	8.1 (3.0)
SlowVAE (LAP-NC)	99.8 (0.6)	95.2 (6.0)	27.6 (8.6)	61.5 (5.3)	96.8 (1.8)	8.4 (3.4)

Table 16: **dSprites.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (first three rows are from (Locatello et al., 2020)). The bottom three rows give mean and standard deviation (s.d.) for the models presented in this paper.

Increasing mean(Δt) leads to increased diversity, and thus more information in the learning signal. However, it is worth noting that since SlowVAE assumes $\alpha = 1$ in the transitions, an increase in α from increasing the temporal gap leads to a reduction in mismatch.

Our results on increasing the temporal difference within pairs of inputs is in agreement with recent work by Oord et al. (2018, Table 2), who show increased performance in representation learning for larger separation between positive samples in a contrastive objective function. Additional related work from Tschannen et al. (2019) shows that temporal separation between frame embeddings influences the representation that is learned from videos.

G.4 Latent Space Visualizations

We visualize differences in learned latent representations using image embedding in Figures 15- 28. We show four different plots for each dataset considered and include all available models. Each figure corresponds to a different dataset.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	100.0	87.9	8.8	22.5	90.2	1.0
Ada-ML-VAE (LOC)	100.0	87.4	14.7	45.6	94.6	2.8
Ada-GVAE (LOC)	100.0	90.2	15.0	54.0	93.9	9.4
SlowVAE (UNI)	100.0 (0.0)	90.4 (0.4)	15.7 (1.5)	48.9 (1.7)	95.7 (1.0)	1.6 (0.4)
SlowVAE (LAP)	100.0 (0.0)	91.0 (2.5)	9.7 (1.1)	51.0 (2.2)	94.4 (1.1)	1.7 (0.9)
SlowVAE (LAP-NC)	100.0 (0.0)	90.8 (1.1)	9.3 (1.1)	50.0 (2.0)	94.6 (0.9)	0.9 (0.9)
SlowVAE (UNI)	100.0 (0.0)	90.4 (0.5)	15.4 (2.2)	48.0 (2.4)	95.4 (1.5)	1.6 (0.5)
SlowVAE (LAP)	100.0 (0.0)	90.2 (3.5)	10.4 (1.8)	50.9 (2.7)	94.1 (1.2)	2.0 (1.1)
SlowVAE (LAP-NC)	100.0 (0.0)	90.9 (1.2)	9.5 (1.4)	50.2 (2.7)	95.0 (1.2)	1.7 (1.4)

Table 17: **Cars3D.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (first three rows are from (Locatello et al., 2020)). The bottom three rows give mean and standard deviation (s.d.) for the models presented in this paper.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	74.0	49.5	21.4	28.0	89.5	9.8
Ada-ML-VAE (LOC)	91.0	72.1	31.1	34.1	86.1	15.3
Ada-GVAE (LOC)	87.9	55.5	25.6	33.8	78.8	10.6
SlowVAE (UNI)	78.8 (2.1)	46.2 (1.9)	23.7 (1.3)	28.8 (0.6)	92.1 (1.6)	7.8 (1.0)
SlowVAE (LAP)	86.0 (0.2)	72.9 (0.7)	25.8 (0.5)	42.7 (0.9)	97.7 (0.3)	6.5 (0.4)
SlowVAE (LAP-NC)	86.1 (0.7)	73.7 (0.6)	26.3 (0.5)	42.5 (0.6)	97.6 (0.3)	6.5 (0.9)
SlowVAE (UNI)	78.2 (3.8)	47.0 (2.9)	23.8 (1.8)	28.7 (0.7)	90.9 (2.1)	7.8 (1.1)
SlowVAE (LAP)	85.9 (0.3)	73.1 (0.9)	25.7 (0.6)	42.6 (0.9)	97.5 (0.3)	6.8 (0.5)
SlowVAE (LAP-NC)	85.7 (1.0)	73.3 (0.8)	26.2 (0.7)	42.6 (0.8)	97.6 (0.5)	6.6 (1.3)

Table 18: **SmallNORB.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (first three rows are from (Locatello et al., 2020)). The bottom three rows give mean and standard deviation (s.d.) for the models presented in this paper.

In Figures 15- 21 we display the mean correlation coefficient matrix and the latent representations for each ground-truth, as described in the main text for Fig. 5.

The top row is the sorted absolute correlation coefficient matrix between the latents (rows) and the ground truth generating factors (columns). The latent dimensions are permuted such that the sum on the diagonal is maximal. This is achieved by an optimal, non-greedy matching process for each ground truth factor with its corresponding latent, as described in appendix C. As such, a more

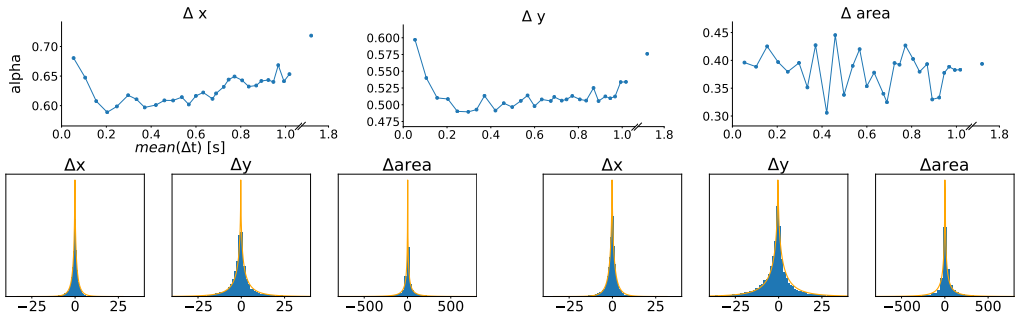


Figure 14: **KITTI Masks Sparseness**. We show the sparseness over time of the transitions for horizontal (Δx), vertical (Δy) as well as mask/object size (Δarea) in KITTI Masks by plotting the α of a generalized Laplace fit for different mean(Δt) (top). To display the quality of the fits, we show two exemplary fits at mean(Δt) $= 0.63$ (bottom-left) and mean(Δt) $= 1.02$ (bottom-right).

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	98.6	83.9	22.0	58.8	93.8	6.2
Ada-ML-VAE (LOC)	100.0	100.0	50.9	94.0	98.8	12.7
Ada-GVAE (LOC)	100.0	100.0	56.2	94.6	97.5	15.3
SlowVAE (UNI)	100.0 (0.1)	97.3 (4.0)	64.4 (8.4)	82.6 (4.4)	95.5 (1.6)	5.8 (0.9)
SlowVAE (LAP)	100.0 (0.0)	95.9 (2.6)	62.5 (3.1)	85.6 (4.0)	98.1 (0.6)	8.2 (1.7)
SlowVAE (LAP-NC)	100.0 (1.6)	97.0 (2.0)	63.6 (5.4)	86.7 (4.1)	98.4 (1.4)	7.0 (2.1)
SlowVAE (UNI)	99.9 (0.3)	95.4 (5.2)	58.8 (13.0)	82.3 (5.4)	95.2 (2.0)	5.7 (1.4)
SlowVAE (LAP)	100.0 (0.0)	95.0 (3.2)	61.5 (4.5)	85.0 (4.7)	98.3 (0.8)	8.9 (2.6)
SlowVAE (LAP-NC)	98.4 (4.9)	97.4 (2.4)	61.6 (10.6)	86.1 (5.2)	98.2 (1.6)	8.2 (2.6)

Table 19: **Shapes3D.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (first three rows are from (Locatello et al., 2020)). The bottom three rows give mean and standard deviation (s.d.) for the models presented in this paper.

Model (Data)	BetaVAE	FactorVAE	MIG	DCI	Modularity	SAP
β-VAE (*i.i.d.*)	54.6	32.2	7.2	19.5	87.4	3.7
Ada-ML-VAE (LOC)	72.6	47.6	24.1	28.5	87.5	7.4
Ada-GVAE (LOC)	78.9	62.1	28.4	40.1	91.6	21.5
SlowVAE (UNI)	58.5 (0.9)	38.6 (2.3)	32.2 (1.0)	29.9 (1.3)	89.2 (2.0)	8.8 (0.8)
SlowVAE (LAP)	67.6 (6.1)	42.4 (6.1)	32.0 (1.8)	35.9 (2.2)	89.5 (1.5)	9.7 (0.8)
SlowVAE (LAP-NC)	60.1 (2.7)	39.2 (1.7)	30.6 (0.7)	34.3 (0.7)	85.9 (1.1)	9.3 (0.9)
SlowVAE (UNI)	58.6 (1.1)	38.5 (3.2)	32.2 (1.2)	30.1 (1.6)	89.4 (2.6)	8.7 (1.0)
SlowVAE (LAP)	66.6 (6.9)	45.5 (8.3)	32.9 (2.6)	35.5 (2.7)	89.2 (1.9)	9.7 (1.2)
SlowVAE (LAP-NC)	61.0 (3.6)	40.3 (2.5)	30.4 (0.8)	34.2 (1.0)	86.6 (1.7)	9.3 (1.0)

Table 20: **MPI3D.** Median and absolute deviation (a.d.) metric scores across 10 random seeds (first three rows are from (Locatello et al., 2020)). The bottom three rows give mean and standard deviation (s.d.) for comparison with other tables.

prevalent diagonal structure corresponds to a better mapping between the ground-truth factors and latent encoding.

The middle set of plots are latent embeddings of random training data samples. The x-axis denotes the ground truth generating factor and the y-axis denotes the corresponding latent factor as matched according to the main diagonal of the correlation matrix. For each dataset, we further color-code the latents by a categorical variable as denoted in each figure.

The bottom set of plots show the ground truth encoding compared to the second best latent as opposed to the diagonally matched latent. This plot can be used to judge how much the correspondence between latents is one-to-one or rather one-to-many.

To further investigate the latent representations, we show a scatter plot over the best and second best latents in figures 22-28. Here, the color-coding is matched by the ground truth factor denoted in each row.

When comparing the correlation matrix with the corresponding scatter plots, one can see that embeddings with sinusoidal curves have low correlation, which illustrates a shortcoming of the metric. Another limitation is that categorical variables which have no natural ordering have an order-dependent MCC score, indicating the permutation variance of MCC. With SlowVAE, we can infer three different types of embeddings. First, we have simple ordered ground truth factors with non-

Model	γ	λ	Data	Permuted?	BetaVAE	FactorVAE	MIG	MCC	DCI	Modularity	SAP
SlowVAE	10	6	Natural (Discrete)	Yes	77.6 (4.1)	69.7 (6.5)	8.5 (4.4)	49.9 (3.5)	17.6 (2.8)	89.8 (3.2)	1.8 (0.9)
SlowVAE	10	6	Natural (Discrete)	No	**82.6** (2.2)	**76.2** (4.8)	11.7 (5.0)	52.6 (4.1)	18.9 (5.5)	88.1 (3.6)	**4.4** (2.3)

Table 21: Impact of removing natural dependence on Discrete Natural Sprites.

Model	γ	λ	Data	Permuted?	MCC
SlowVAE	10	6	Natural (Continuous)	Yes	52.9 (4.2)
SlowVAE	10	6	Natural (Continuous)	No	49.1 (4.0)

Table 22: Impact of removing natural dependence on Continuous Natural Sprites.

circular boundary conditions. Here, SlowVAE models often show a clear one-to-one correspondence (e.g. Fig 22 scale, x-position and y-position; Fig 25 θ-rotation; Fig 26 Φ-rotation). Second, we observe circular embeddings due to boundary conditions for certain factors (e.g. Fig 15, 22 3rd row; Fig 16, 23 2nd row). Note that not all datasets with orientations exhibit full rotations and thus do not have circular boundary conditions, e.g. smallNORB. Finally, we have categorical variables, where no order exists (e.g. Fig. 16, 23 top row, Fig 17, 24 top row, Fig 18, 25 top row) resulting in separated but not necessarily ordered clusters.

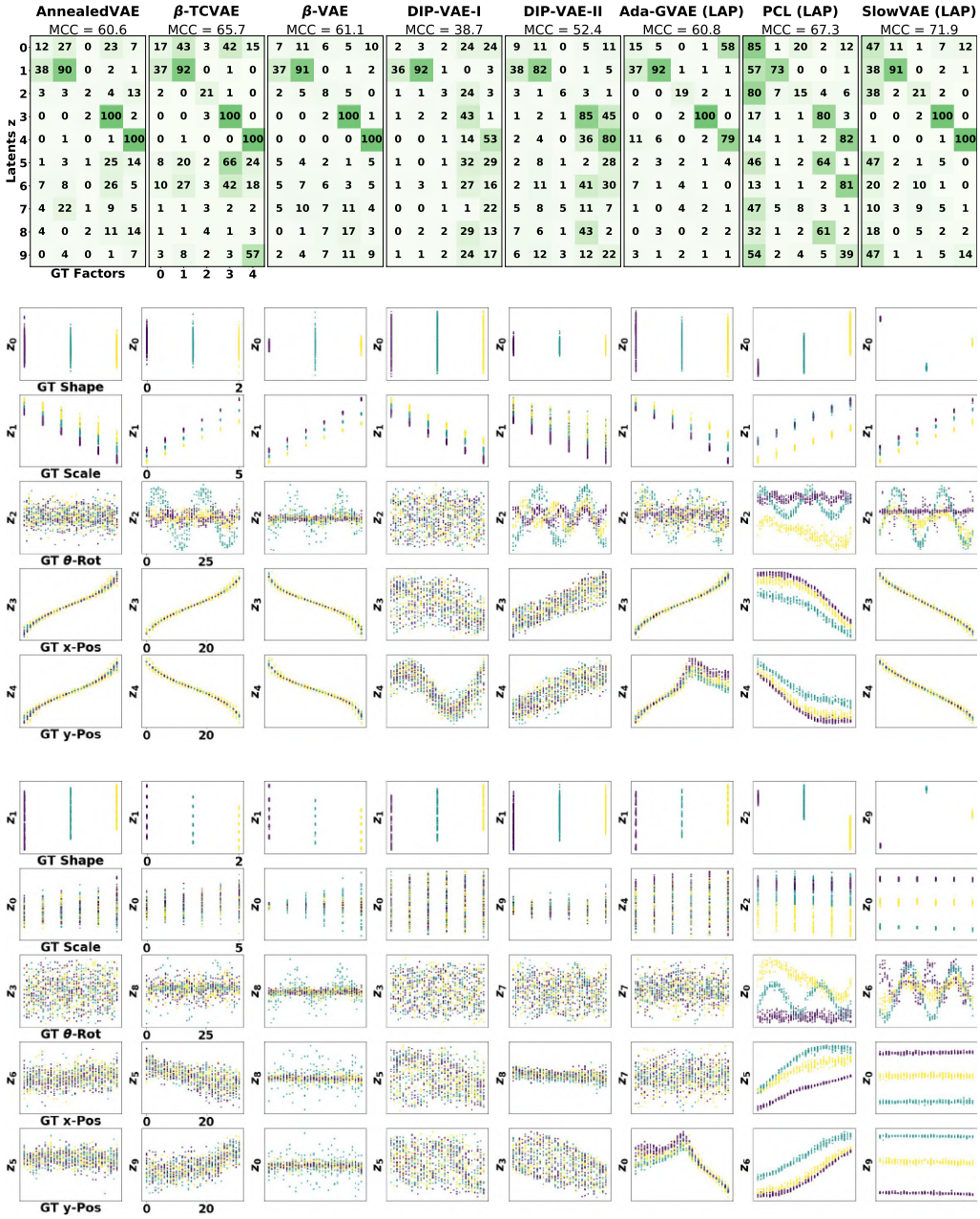


Figure 15: **DSprites Latent Representations.** Top, MCC correlation matrices. Middle five rows, model latent over highest correlating ground truth factor. Bottom five rows, model latent over second highest correlating ground truth factor. The color-coding corresponds to the shapes: heart/yellow, ellipse/turquoise and square/purple.

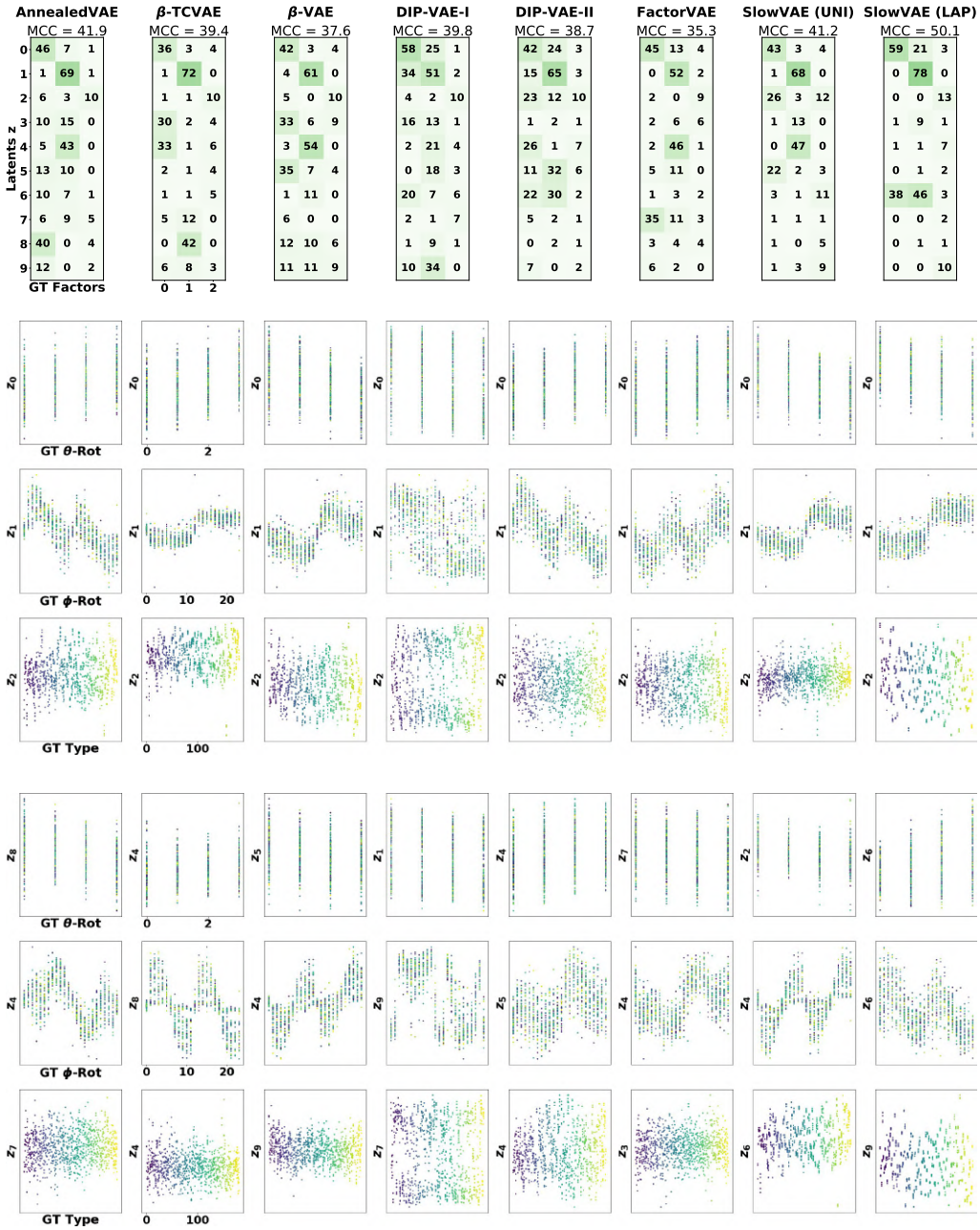


Figure 16: **Cars3D Latent Representations.** Top, MCC correlation matrices. Middle three rows, model latent over highest correlating ground truth factor. Bottom three rows, model latent over second highest correlating ground truth factor. The color-coding corresponds to the 183 different car types (GT Types) in the dataset.

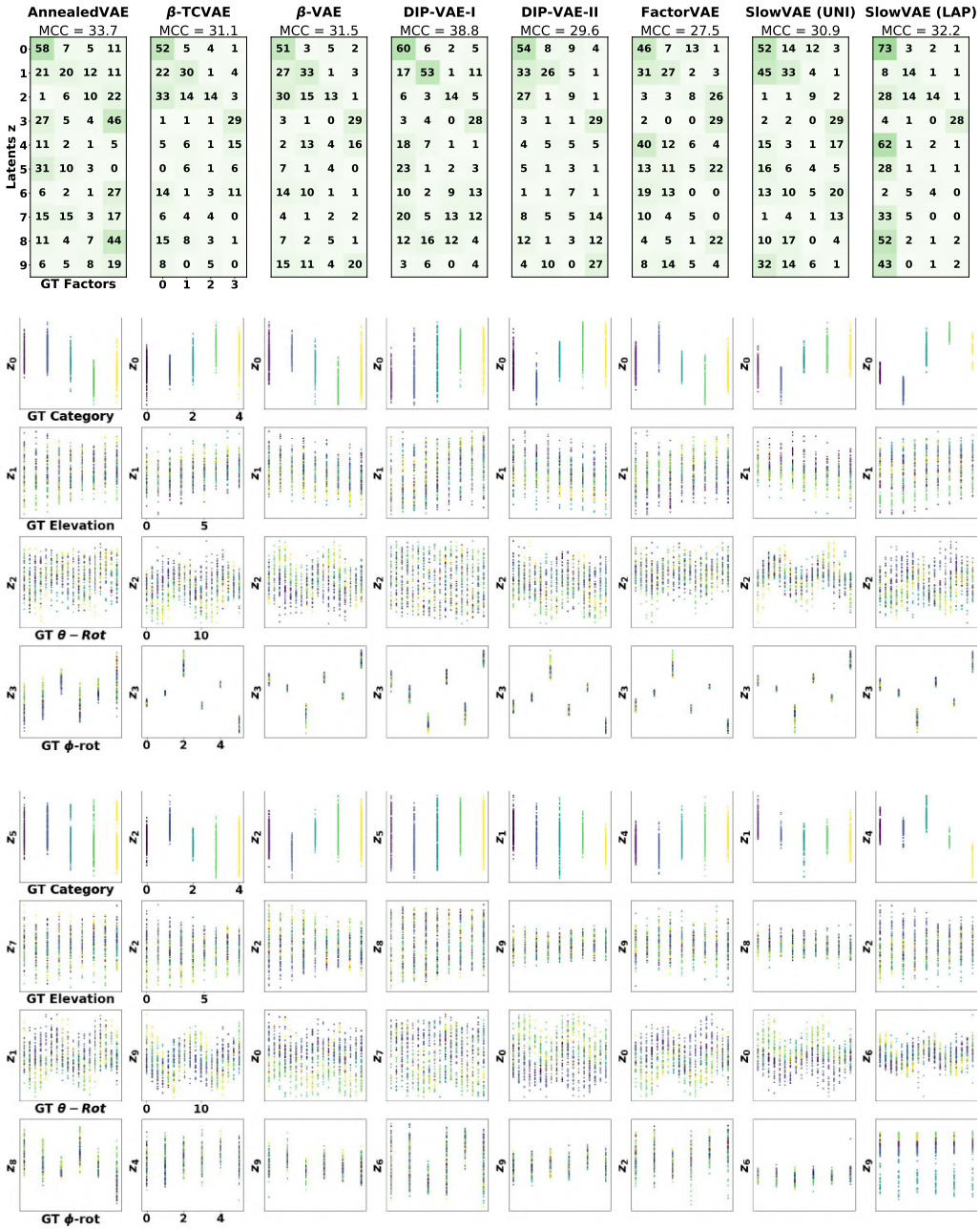


Figure 17: **SmallNorb Latent Representations.** Top, MCC correlation matrices. Middle four rows, model latent over highest correlating ground truth factor. Bottom four rows, model latent over second highest correlating ground truth factor. The color-coding corresponds to the five different GT categories in the dataset.

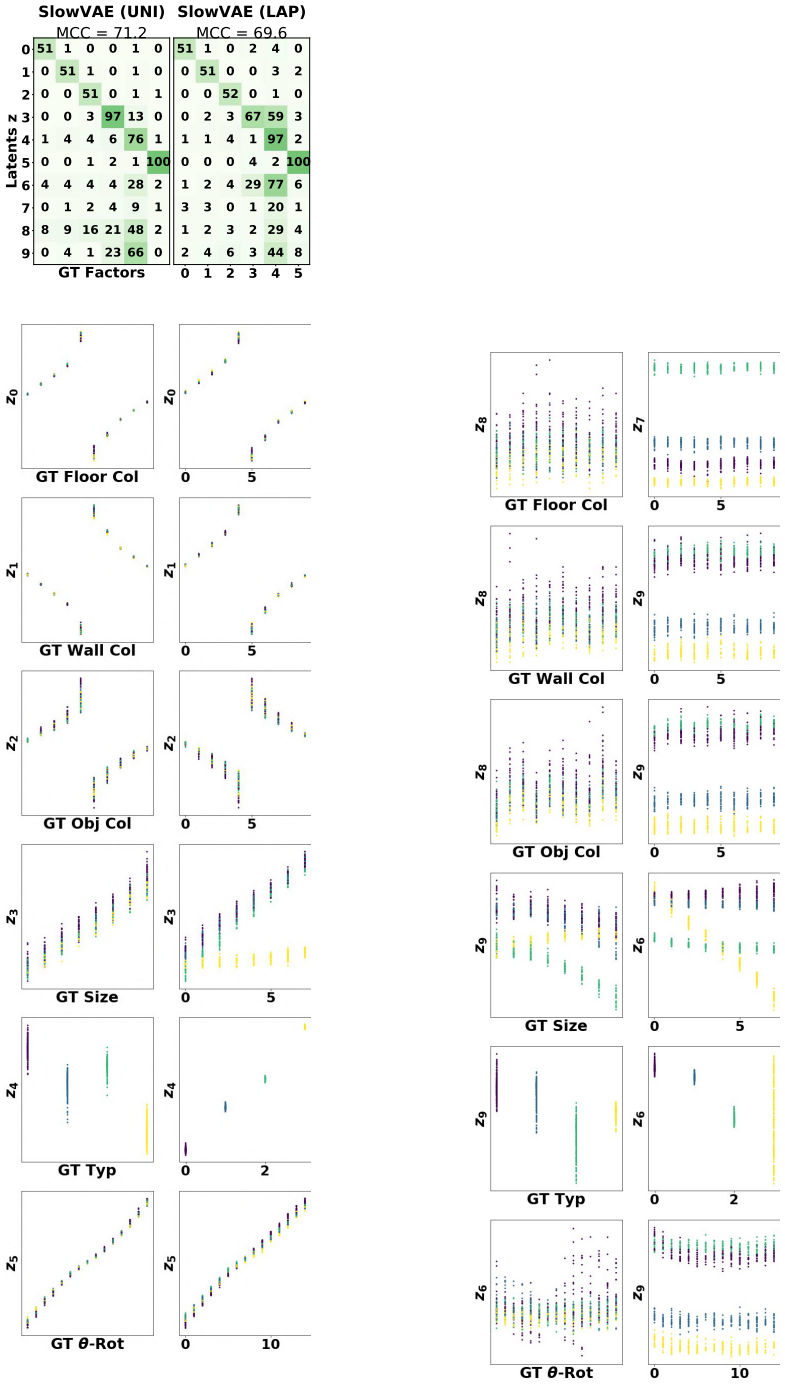


Figure 18: **Shapes3D Latent Representations.** Top, MCC correlation matrices. Left two columns, model latent over highest correlating ground truth factor. Right two columns, model latent over second highest correlating ground truth factor. The color-coding corresponds to the four different object types (GT-Type) in the dataset.

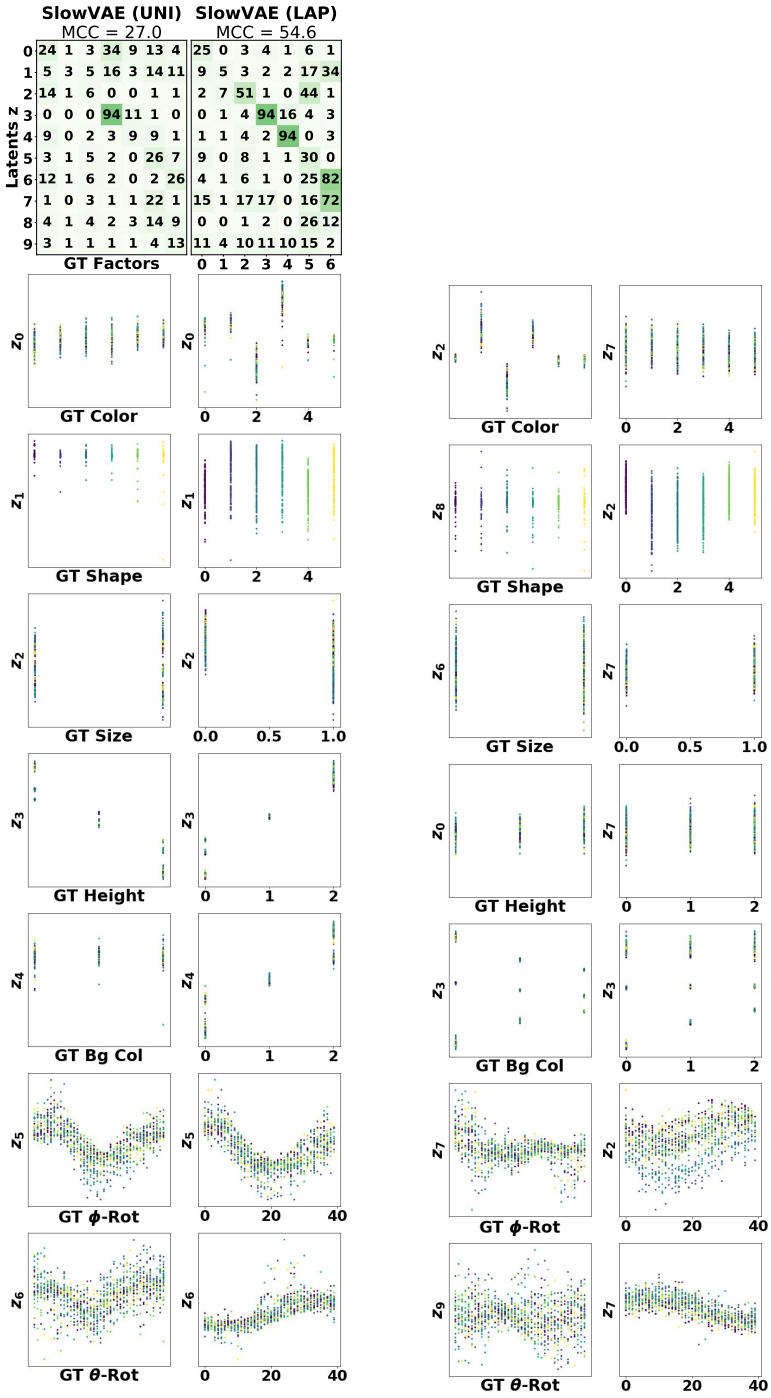


Figure 19: **MPI3DReal Latent Representations.** Top, MCC correlation matrices. Left two columns, model latent over highest correlating ground truth factor. Right two columns, model latent over second highest correlating ground truth factor. The color-coding corresponds to the six different object shapes (GT Shape) in the dataset.

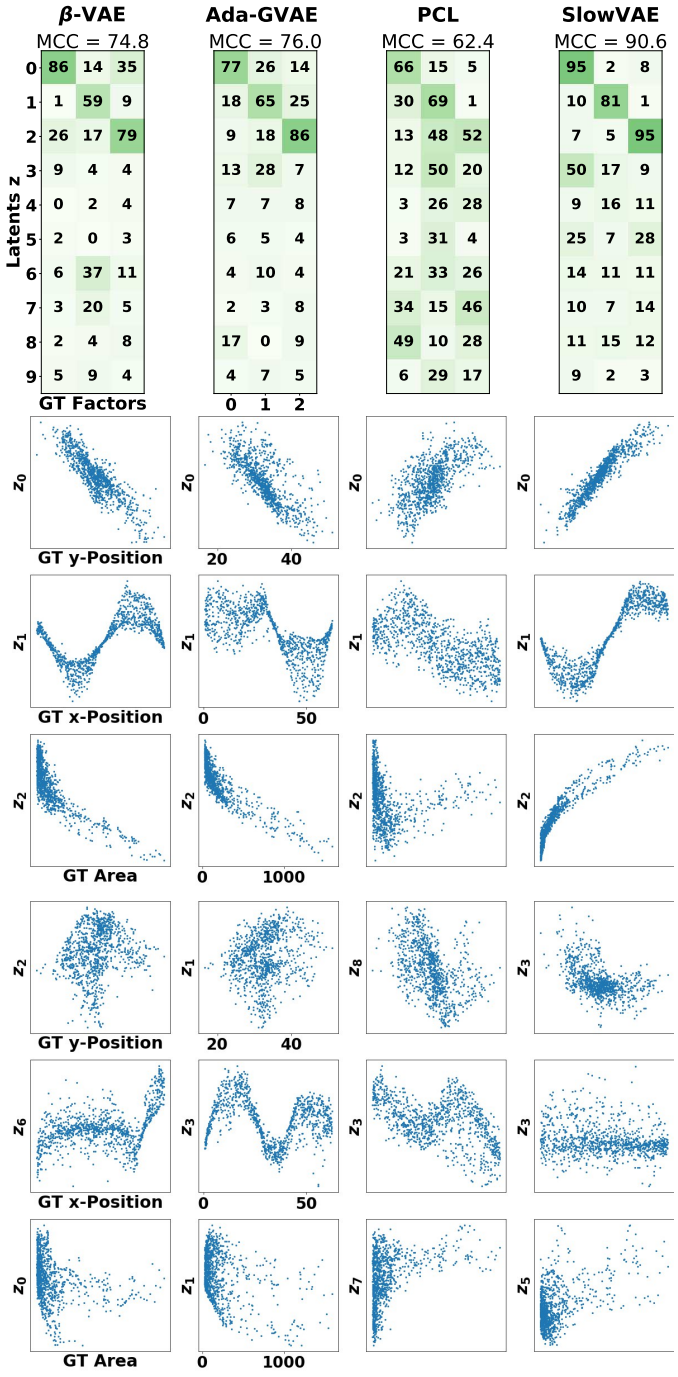


Figure 20: **KITTI Masks Latent Representations.** Top, MCC correlation matrices. Middle three rows, model latent over highest correlating ground truth factor. Bottom three rows, model latent over second highest correlating ground truth factor.

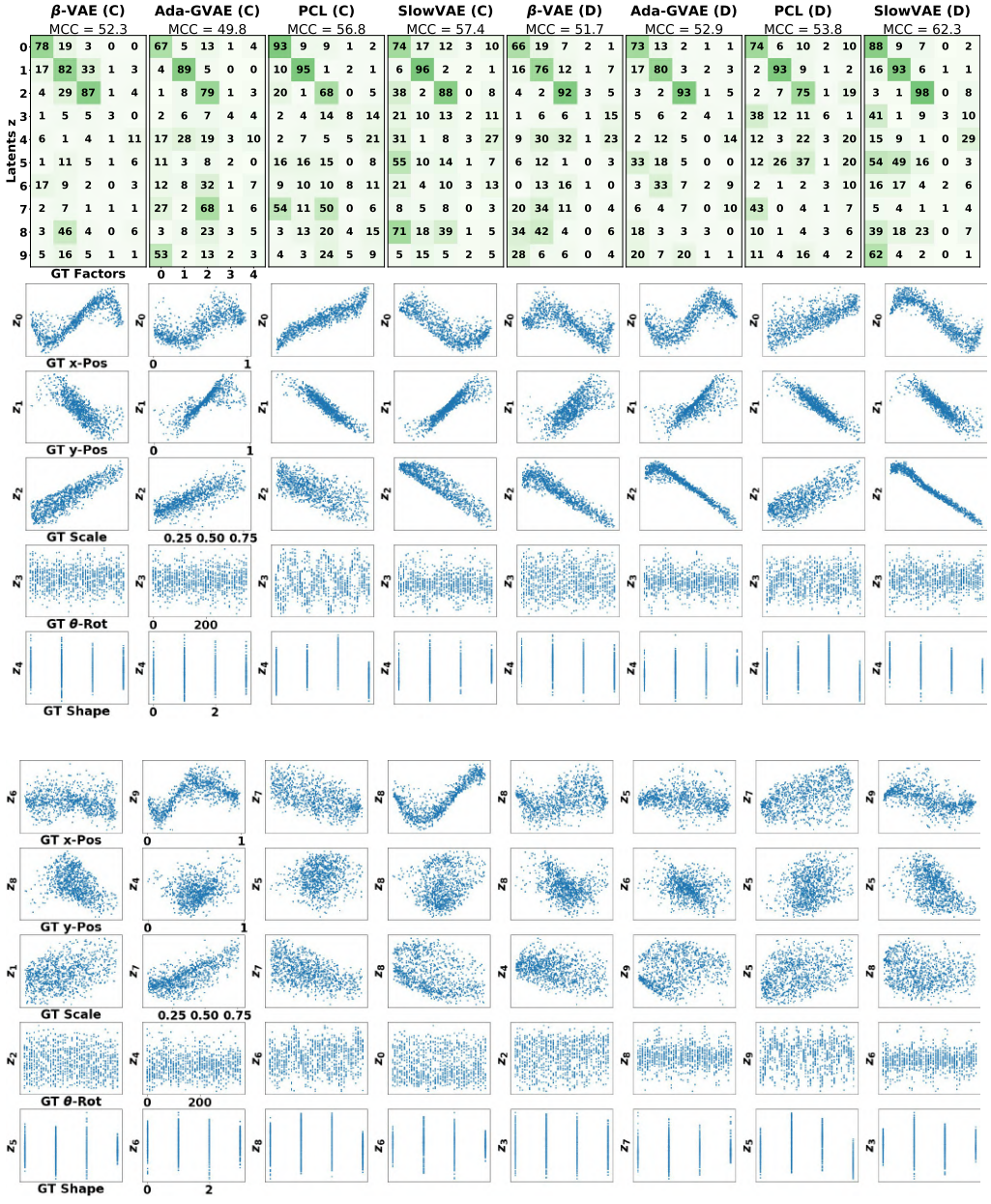


Figure 21: **Natural Sprites Latent Representations.** Top, MCC correlation matrices. Middle five rows, model latent over highest correlating ground truth factor (colored by category). Bottom five rows, model latent over second highest correlating ground truth factor. The left two columns denote the continuous (C) version of Natural Sprites, whereas the right two columns correspond to the discretized (D) version.

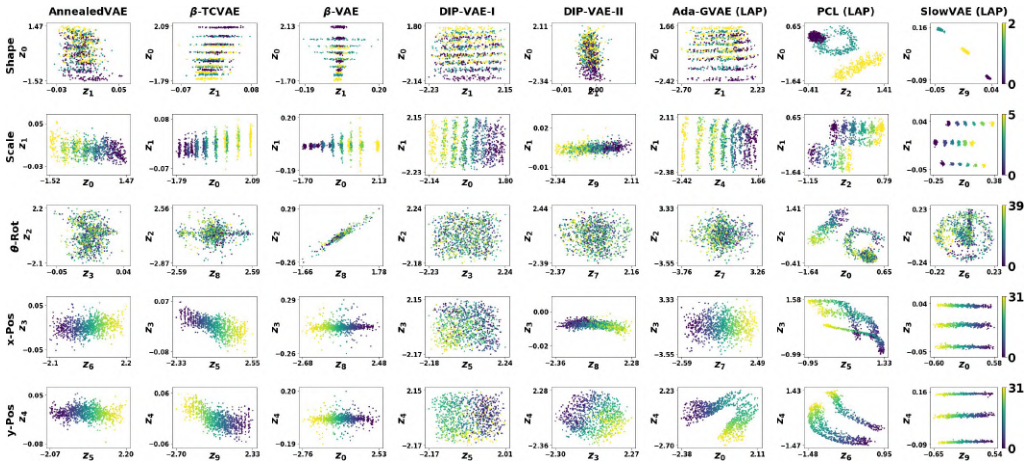


Figure 22: **DSprites Latent Representations**. Best two latents selected from Fig 15. Color-coded by the corresponding ground truth factor.

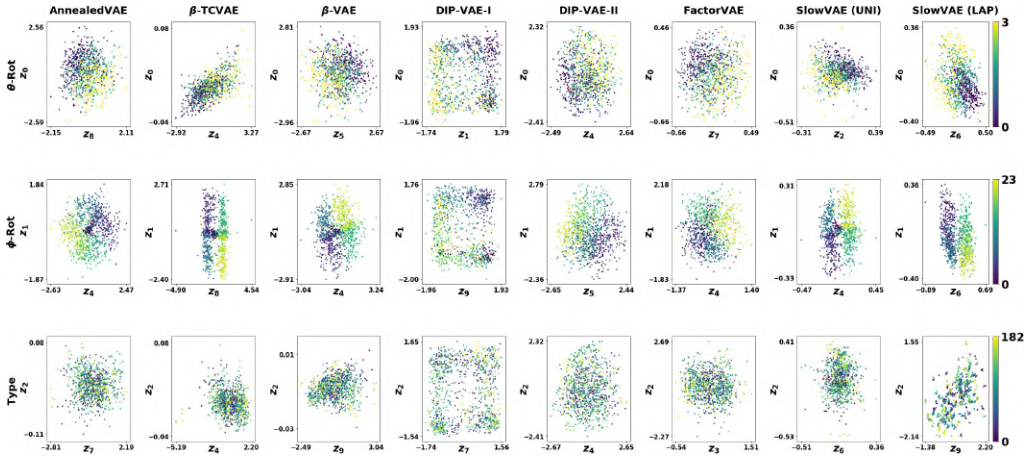


Figure 23: **Cars3D Latent Representations**. Best two latents selected from Fig 16. Color-coded by ground truth.

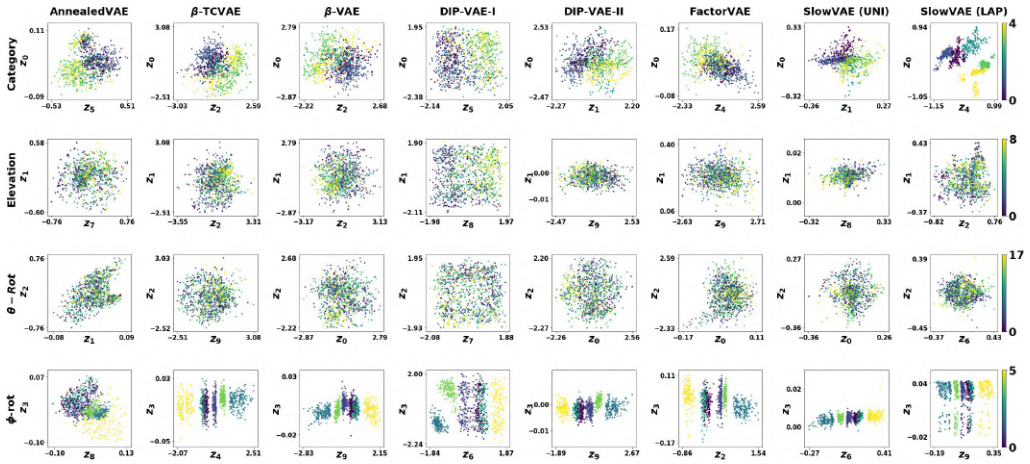


Figure 24: **SmallNorb Latent Representations**. Best two latents selected from Fig 17. Color-coded by ground truth.

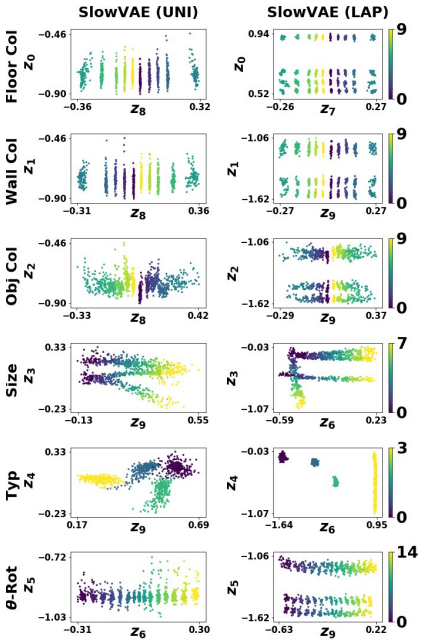


Figure 25: **Shapes3D Latent Representations**. Best two latents selected from Fig 18. Color-coded by ground truth.

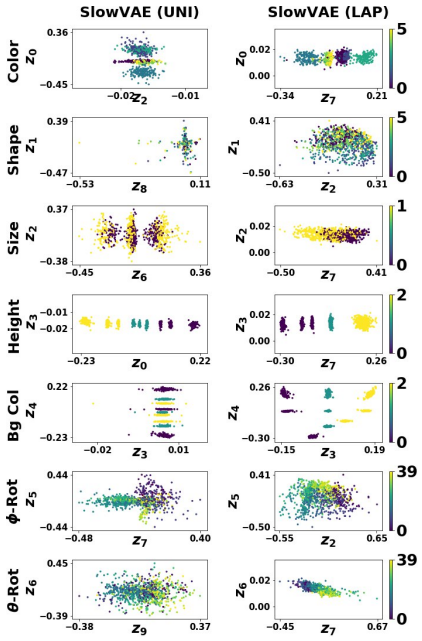


Figure 26: **MPI3DReal Latent Representations**. Best two latents selected from Fig 19. Color-coded by ground truth.

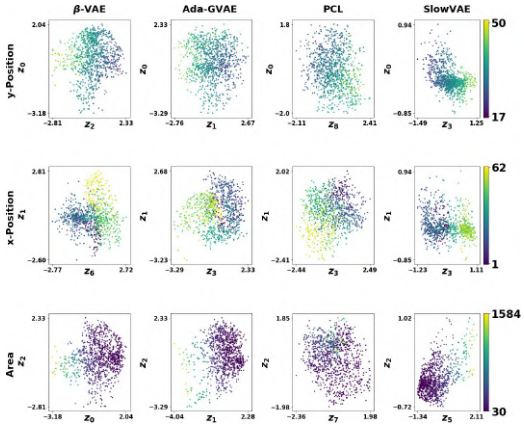


Figure 27: **KITTI Masks Latent Representations**. Best two latents selected from Fig 20. Color-coded by ground truth.

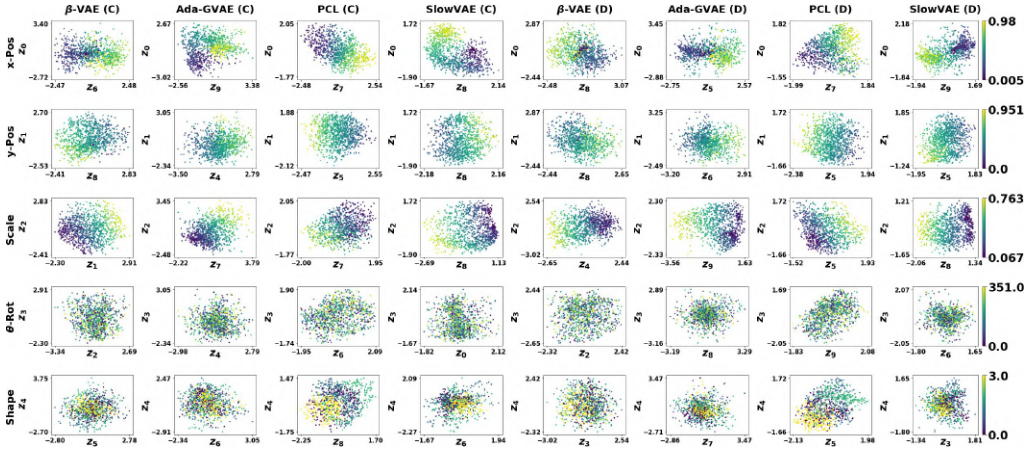


Figure 28: **Natural Sprites Latent Representations**. Best two latents selected from Fig 21. The left four columns denote the continuous (C) version of Natural Sprites, whereas the right four columns correspond to the discretized (D) version. Color-coded by ground truth.

Discussion 3

While significant time has passed since (Klindt et al., 2021), disentangled representation learning still hasn't achieved real-world impact in practice. We will discuss recent research pertaining to this topic, clarify what still remains unresolved, and propose approaches for moving past these long-standing problems.

3.1 Why annotate factors of variation?

What is the state of progress? In (Zimmermann et al., 2021), we saw that state-of-the-art approaches to visual SSL can yield disentangled representations if certain statistical assumptions are fulfilled, most notably that positive pairs are constructed such that all FoVs are able to change. Thus, by enforcing assumptions on the distribution of the latent sources, i.e. the FoVs, disentanglement is achieved. Assumptions on the distribution of latent sources is what enabled the first successful methodology for nonlinear ICA (Hyvarinen and Morioka, 2016, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020a; Ahuja et al., 2022b), i.e. identifying the latent sources from observational data when the relationship between the latent sources and the observational data is nonlinear. These assumptions were necessary since, without additional structure, nonlinear ICA has been proven to be impossible (Darmois, 1953; Hyvärinen and Pajunen, 1999; Locatello et al., 2019). This breakthrough in nonlinear ICA set the stage for research exploring the space of possible assumptions that enable identifying the latent sources, thereby providing a structured direction for understanding when representations disentangle.

A number of works have furthered this direction of assuming structure on the distribution of latents for identifiability. Lachapelle et al. (2022, 2024) takes a step away from assuming statistical independence (which, as we saw in Figure 1.3 (right), can be violated in practice) by considering the setting where a change in FoV can (temporally) depend on other FoVs, and even on an action, e.g. an intervention (Locatello et al., 2020a; Ahuja et al., 2022a; Lippe et al., 2022, 2023b; Squires et al., 2023; Zhang et al., 2023; Ahuja et al., 2023a; Jiang and Aragam, 2023; von Kügelgen et al., 2023; Buchholz et al., 2023; Sturma et al., 2023; Morioka and Hyvärinen, 2023; Bing et al., 2023; Ahuja et al., 2023b; Zhang et al., 2024a; Varıcı et al., 2024), showing that disentanglement here hinges on the sparsity of this dependency graph. (Lippe et al., 2023a) instead shows that constraining actions to be binary toggle switches for interventions on FoVs is sufficient for disentanglement, empirically showing promising results for robotics applications (Kolve et al., 2017). Lastly, Brehmer et al. (2022) took this a step further by considering FoVs to depend on other FoVs according to a structural causal model (SCM) (Pearl, 2009; Peters et al., 2017; Schölkopf et al., 2021) and demonstrates that both the FoVs can be disentangled and the causal graph can be recovered if atomic, perfect interventions on the FoVs are available, i.e. counterfactuals (Locatello et al., 2020a; Von Kügelgen et al., 2021; Ahuja et al., 2022a). Follow-up work (Squires et al., 2023; Ahuja et al., 2023a; von Kügelgen et al., 2023) has relaxed these assumptions by shifting from a multi-view to a multi-environment setting, thereby considering unpaired samples from interventional distributions.

Recently, another direction has emerged for assuming structure that enables identifiability. So far, all of the aforementioned research has assumed structure on the FoVs, namely how they

are sampled, but none has assumed structure on the nonlinear *mixing* that maps the FoVs to the observational data. (Gresele et al., 2021) proposed independent mechanism analysis (IMA), formalized as an orthogonality condition on the columns of the Jacobian of the mixing function, by applying the principle of independent causal mechanisms (Janzing and Schölkopf, 2010; Schölkopf et al., 2012; Peters et al., 2017), which postulates that the generative process consists of independent modules which do not share information. IMA has been used to characterize how the variational autoencoder (VAE) (Kingma and Welling, 2013) objective favors decoders with a column-orthogonal Jacobian (Rolinek et al., 2019; Kumar and Poole, 2020; Reizinger et al., 2022). However, while IMA does rule out well-known counterexamples to identifiability (Gresele et al., 2021; Ghosh et al., 2023), a proof of identifiability remains an open question (Buchholz et al., 2022). On the other hand, (Moran et al., 2022) proves identifiability under a fixed sparsity structure where each feature of the observational data depends only on a small subset of the FoVs. Similarly, (Zheng et al., 2022) prove identifiability assuming that, for each FoV, there exists a set of observed feature(s) that is only dependent on said FoV. (Brady et al., 2023) showed that assuming the mixing is *compositional* (Fodor and Pylyshyn, 1988), i.e. each observed feature depends on at most one latent slot, enables proving identifiability for object-centric learning (Burgess et al., 2019; Locatello et al., 2020b; Greff et al., 2020), thus showing that the representation disentangles objects (Higgins et al., 2018; Wulfmeier et al., 2021). Finally, (Lachapelle et al., 2023a) shows that, as long as a compositional mixing function (Brady et al., 2023) is twice continuously differentiable, it is additive, i.e. the function consists of a summation after each slot is decoded independently, which is equivalent to having a (block) diagonal Hessian (Peebles et al., 2020), and lends itself to compositional generalization (Sauer and Geiger, 2021; Montero et al., 2022; Wiedemer et al., 2023), where the learned decoder can extrapolate to unseen combinations of seen objects.

What open problems remain? The impossibility of identifiability (Darmois, 1953; Hyvärinen and Pajunen, 1999; Locatello et al., 2019) has encouraged an extensive search for additional structure to impose on the problem for bypassing the impossibility result. Without additional structure, counterexamples such as the Darmois construction (Darmois, 1953; Hyvärinen and Pajunen, 1999) demonstrate that one can infer latent variables that are statistically indistinguishable from the ground-truth latent sources, but are entirely entangled w.r.t. the latent sources. Thus, we require additional structure to obtain reliable learning signal for recovering the latent sources. While there isn't evidence to suggest that the counterexamples used to prove impossibility occur in practice, for identifiability, we still have to show that we have sufficient structure in practice to rule out all counterexamples, and therefore bypass the impossibility result. However, since the identifiability problem reduces to inverting the data generating process (Zimmermann et al., 2021), but the data generating process is, for the most part, unknown in practice, how can we justify that the additional structure assumed for identifiability actually holds in practice? Furthermore, in practice, how can we measure performance, given we do not know the latent sources that underly the data generating process?

While we do advocate for further efforts towards validating assumptions in practice, going beyond what was done in (Klindt et al., 2021) may not be straightforward. There, we required annotated masks (Geiger et al., 2012; Milan et al., 2016; Xu et al., 2018; Yang et al., 2019; Voigtlaender et al., 2019) to estimate object scale and object position, which are of course only 2 variables we are considering members of the unknowable set of FoVs. Thus, the same data labeling bottleneck that motivates SSL (Balestriero et al., 2023) is in effect when attempting to validate assumptions or evaluate performance for disentanglement in practice. Hence, we do not have a scalable approach to validating assumptions for more FoVs in practice, which is problematic for disentanglement. While state-of-the-art models may be reliable enough to

automate annotation (Radford et al., 2021; Kirillov et al., 2023; Huang et al., 2023), if we can only disentangle in practice what we have state-of-the-art models for, the practical utility of disentanglement would be unclear.

With that said, as an alternative to photorealistic simulation (Bordes et al., 2023b) and synthetic data generation (Wiles et al., 2022), ImageNet-X (Idrissi et al., 2023) provides human annotated object FoVs for the ImageNet (Russakovsky et al., 2015) validation set and a random subset of training images. While none of the FoVs are continuous, statistical dependencies between FoVs could still be found, yielding a dependency graph between FoVs which may provide sufficient structure for identifiability. Furthermore, given data augmentations used in practice by state-of-the-art visual SSL approaches can already be interpreted as counterfactuals in the underlying latent SCM (Von Kügelgen et al., 2021; Pearl, 2009), it would be interesting to consider if atomic, perfect interventions could indeed be possible for the given FoVs (Brehmer et al., 2022). An inferred dependency graph may imply a partitioning of the dataset into multiple environments (Creager et al., 2021), which enables testing methodology that assumes unpaired samples from interventional distributions (Squires et al., 2023; Ahuja et al., 2023a; von Kügelgen et al., 2023). It is possible that structure could be uncovered in the relationship between the pixels and the annotated FoVs, which may lead to a direction for assessing structure assumed on the mixing function in practice (Moran et al., 2022; Zheng et al., 2022; Brady et al., 2023). Finally, given any auxiliary information (Hyvarinen et al., 2019) here would have to be imposed (Von Kügelgen et al., 2021) or inferred (Creager et al., 2021), disentanglement approaches that do not assume additional auxiliary information (Higgins et al., 2017a; Locatello et al., 2019) would be a natural fit, for which there is promising recent work worth considering (Wang and Jordan, 2021; Kivva et al., 2022; Funke et al., 2022; Roth et al., 2023; Hsu et al., 2023).

What is the way forward? Considering how the numerous identifiable methods for disentanglement could be applied to disentangle the FoVs in ImageNet-X (Idrissi et al., 2023) is a good example of an exercise we should be pursuing when we consider additional structure for identifiability. Without any exploratory data analysis to find additional structure, e.g. dependencies, to exploit, a straightforward approach would be to leverage atomic, perfect interventions on the FoVs, i.e. counterfactuals (Von Kügelgen et al., 2021; Ahuja et al., 2022a; Brehmer et al., 2022). As shown in (Eastwood et al., 2023), if, for each FoV, we have augmentations that uniquely leave said FoV invariant, or unchanged, we can disentangle the FoVs. However, can we really assume counterfactuals for each FoV are available in practice?

In the supervised setting, i.e. we train on ImageNet-X, and thus can leverage the FoV annotations for learning, for each FoV, we can simply construct positive pairs which only have said FoV in common (Khosla et al., 2020). However, this approach is bottlenecked by data labels, and, more importantly, identifiable methods for disentanglement aren't needed in the supervised setting. In the self-supervised setting, i.e. we only use ImageNet-X for validation/evaluation, note that, for each FoV, the only constraint on the positive pair is to have said FoV in common. Therefore, constructing positive pairs by independently sampling from a generative model conditioned on the FoV of interest could be a solution (Wiles et al., 2022), especially since recent work suggests that learning from synthetic data is starting to rival learning from real data (Azizi et al., 2023; Tian et al., 2023b; Fan et al., 2023; Tian et al., 2023a; Afkanpour et al., 2024).

As discussed, having disentanglement in practice be dependent on the performance of state-of-the-art models can be a dissatisfying solution, particularly since the same approach could enable simply supervised learning on purely synthetic data. Thus, in order to obtain a reliable method for self-supervised disentanglement in practice, we require exploratory data analysis to obtain additional structure. In all, we have seen many examples of additional structure enabling reliable self-supervised disentanglement, what we're missing is additional structure that has been

validated on the real-world data we would like to disentangle in practice. For that, research contributions like ImageNet-X (Idrissi et al., 2023) provide an actionable way forward.

3.2 Why disentangled representation learning?

What is the state of progress? In (Klindt et al., 2021), we saw that, in video, FoVs of interest, e.g. position and scale, change sparsely over time, and we showed that leveraging that structure can enable theoretically and empirically achieving disentanglement. Notably, this step towards validating assumptions in real-world data was novel relative to prior work (Hyvarinen and Morioka, 2016, 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020a), which also leveraged temporal structure to recover the latent sources, but did not validate their assumptions in practice. For instance, in (Hyvarinen and Morioka, 2017), the second derivative of the probability density function for the conditional distribution is assumed to be less than zero, which is only met for $\alpha > 1$ for generalized Laplace transition distributions, but in real-world temporal data, $\alpha < 1$ for the FoVs studied (see Tab. 6 in (Klindt et al., 2021)). Furthermore, in (Khemakhem et al., 2020a), since the conditional posterior is assumed to be exponential family (Wainwright et al., 2008), but the exponential family only includes Laplacian distributions with fixed mean, the conditional that fit the real-world temporal data does not satisfy the assumption. In all, without validating assumptions in practice, we have no reason to expect that disentanglement methods can be applied in practice, thus we continue to advocate for further efforts in that direction.

With that said, in (Klindt et al., 2021), we did not empirically demonstrate disentanglement on real-world video data. As we computed FoVs from object mask annotations provided by video datasets (Geiger et al., 2012; Milan et al., 2016; Xu et al., 2018; Yang et al., 2019; Voigtlaender et al., 2019), we used these FoVs to extend existing benchmarks (Matthey et al., 2017; Watters et al., 2019; Klindt et al., 2021), and used these masks to create a novel benchmark (Klindt et al., 2021), but we did not benchmark on the real-world videos themselves. Given contrastive learning had already been shown to scale well to complex natural image data (Chen et al., 2020), in (Zimmermann et al., 2021), we scaled up the evaluation to natural image resolution (Deng et al., 2009), but as we do not have the FoVs for natural images, we rendered a dataset using Blender (Blender Online Community, 2021), and did the same for evaluation in (Von Kügelgen et al., 2021). While these contributions have enabled subsequent work (Lippe et al., 2022; Brehmer et al., 2022; Daunhawer et al., 2023; Xu et al., 2024; Talon et al., 2024), in order to demonstrate disentanglement in practice, we must depart from only evaluating disentanglement when all FoVs are known (Klindt et al., 2021).

Since requiring knowledge of all FoVs restricts disentanglement evaluation to highly controlled domains (Gondal et al., 2019; Dittadi et al., 2021), in the absence of FoV annotations (Idrissi et al., 2023), what can we do instead? Of course, human inspection of the disentangled components is always an option (Higgins et al., 2017a; Bordes et al., 2022; Liu et al., 2023a), which is an understandable direction given the desire that a disentangled representation will be interpretable, and is an evaluation that is especially encouraged in scientific applications for uncovering insight (Lopez et al., 2023; Bereket and Karaletsos, 2023; Klindt et al., 2024). Given disentanglement has long been seen as important for achieving robustness to distribution shift (Schölkopf et al., 2021), out-of-distribution (OOD) accuracy has also been used as a proxy for evaluating disentanglement in real settings (Liu et al., 2023b; Fumero et al., 2023; Lachapelle et al., 2023b), particularly in sim2real transfer (Higgins et al., 2017b; Gondal et al., 2019; Dittadi et al., 2021; Liu et al., 2023c). Finally, for disentangling objects, evaluation has consisted of object instance segmentation (Hubert and Arabie, 1985; Greff et al., 2019) using progressively more realistic datasets (Kabra et al., 2021; Kipf et al., 2022; Greff et al., 2022). Recently, progress has

led to scaling to real-world data (Elsayed et al., 2022; Biza et al., 2023; Seitzer et al., 2023; Löwe et al., 2023). Notably, the latest work (Zadaianchuk et al., 2023b; Qian et al., 2023; Aydemir et al., 2023) has demonstrated the ability to scale to YouTube-VIS (Yang et al., 2019), natural video where, in (Klindt et al., 2021), we found position and scale to change sparsely over time.

What open problems remain? Does object instance segmentation imply disentanglement of objects? No, in general, successful object segmentation does not require an underlying representation decomposed into objects (Higgins et al., 2018; Wulfmeier et al., 2021). With that said, if the learned decoder was additive (Brady et al., 2023; Lachapelle et al., 2023a), the decoder output for each slot would be used for segmentation, and here, it would be possible to show that successful segmentation is accompanied by disentanglement (Higgins et al., 2018; Wulfmeier et al., 2021), i.e. slots decompose into objects (Brady et al., 2023; Lachapelle et al., 2023a). Interestingly, while more flexible transformer decoders have been proposed (Vaswani et al., 2017; Singh et al., 2022a,b; Sajjadi et al., 2022; Chang et al., 2022), evidence has shown (Seitzer et al., 2023) that, at segmenting real-world data (Lin et al., 2014), transformer decoders can be outperformed by masked decoders (Burgess et al., 2019; Greff et al., 2019; Locatello et al., 2020b), an extension on additive decoders (Lachapelle et al., 2023a) where masks are produced for each slot, are normalized across slots, and are then used to weight the additive decoder sum over independently decoded slots. While slot-wise masks do facilitate modeling occluded objects, the normalization across slots violates additivity (Lachapelle et al., 2023a), hence additive decoders must be treated as an approximation to decoders used in practice. Still, the resemblance between what's used in practice and what's proven in theory is noteworthy.

However, while autoencoders are still used at scale (Seitzer et al., 2023), to scale up, recent work has shifted from reconstructing pixels to reconstructing features from self-supervised pretraining (Seitzer et al., 2023; Zadaianchuk et al., 2023b; Aydemir et al., 2023). Specifically, features pretrained with non-contrastive SSL (Caron et al., 2021) are used, for which, our evidence (Von Kügelgen et al., 2021) suggests learned representations discard what's variant to data augmentation, and, for data augmentations chosen in practice, position, hue and rotation information are discarded. If so, how could the decoder masks demonstrate object instance segmentation? In (Von Kügelgen et al., 2021), our results rely upon the augmentations leaving certain FoVs, e.g. object class, unchanged, or invariant, thus yielding a shortcut (Geirhos et al., 2020; D'Amour et al., 2022) solution to the SSL problem. However, small, or local, random crops (Szegedy et al., 2015), can lead to a scenario where all FoVs, including object class, are variant, removing the shortcut(s) (Minderer et al., 2020; Robinson et al., 2021), and possibly shifting the setting to be more in line with (Zimmermann et al., 2021), where all information is preserved as identifiability is yielded. Notably, (Caron et al., 2021) use multi-crop (Caron et al., 2020), which uses many local crops as augmentation, thus the likelihood of shortcut removal is increased in their case. Furthermore, our results in (Von Kügelgen et al., 2021) only apply to the representation on which the loss is applied, which for (Caron et al., 2021), is the projection (Chen et al., 2020) of the representation for the CLS token (Devlin et al., 2018; Dosovitskiy et al., 2021), an extra learnable token appended to the sequence of patches the Vision Transformer (ViT) (Dosovitskiy et al., 2021; Vaswani et al., 2017) takes as input, but (Seitzer et al., 2023; Zadaianchuk et al., 2023b) use the representations of the sequence of patches instead. While it remains unclear what information the patch token representations represent, it's been consistently observed that pre-projector representations contain significantly more information than post-projector representations (Chen et al., 2020; Von Kügelgen et al., 2021; Jing et al., 2022; Bordes et al., 2023a; Eastwood et al., 2023). Empirically, while it has been shown that CLS token representations yield impressive foreground segmentation (Caron et al., 2021; Hamilton et al., 2022; Zadaianchuk et al., 2023a), the patch representations provide

dense features (Sundaram et al., 2010) by describing all areas of the image, and have been shown to yield competitive video instance segmentation (Jabri et al., 2020; Pont-Tuset et al., 2017) without any additional training or finetuning (Caron et al., 2021). In all, it is conceivable that dense feature representations of the data via non-contrastive SSL preserve sufficient structure for successful segmentation to be accompanied by object disentanglement.

Lastly, does the latest in video object-centric learning suggest a bias towards disentanglement? (Zadaianchuk et al., 2023b) introduce a temporal similarity prediction task where, for each image patch of a frame at time t, the decoder must map the slots to the semantic similarity, i.e. the cosine similarity, between SSL representations of the patch and each patch k frames in the future. Given the model has been provided with the semantics for each patch (Caron et al., 2021), solving the prediction task requires the slots to represent some notion of what could change and how could it change, which suggests a bias towards object decomposition, since what moves in videos are objects (Tangemann et al., 2023). Note that, assuming the semantics represent FoVs, e.g. position and scale, there may be a connection between this objective and the temporal sparse coding objective in (Klindt et al., 2021) treated as a discriminative learning task (Zimmermann et al., 2021), where the latents at frame t are fed through a nonlinear projection head (Chen et al., 2020), the decoder in this case, and the output is compared to the encoder input at frame $t + k$. Note that the encoder input is a semantic representation (Caron et al., 2021). From this perspective, in order to improve the disentanglement of latents within each slot, given the task is to predict the cosine similarity in representation space, it would be interesting to investigate whether a von Mises-Fisher (vMF) distribution (Zimmermann et al., 2021) does indeed provide the best fit for how the FoVs in the dense semantic representation change over time in YouTube-VIS (Yang et al., 2019).

What is the way forward? Given our focus on vision, it's prudent to ask what more we may want from disentanglement in practice if the recent literature suggests that we already have performant unsupervised video instance segmentation in practice. As discussed, there are routes to showing that instance segmentation comes with the representation decomposing into objects, i.e. object disentanglement (Higgins et al., 2018; Wulfmeier et al., 2021). From there, it may be possible to also disentangle the FoVs, but what exactly are we working towards?

We posit that the practical utility of disentanglement in vision comes in the form of planning. While autonomous driving is a safety-critical scenario for which reliable decision-making is far from solved, the domain is constrained enough for there to be fairly simple, universal reasoning traces for decision-making. For example, if the perception module has detected that an object, e.g. an intersection, stop sign, or a person, is approaching, decelerate quickly enough such that the vehicle state will change to stopped well before any contact could conceivably occur, wait for a reasonable amount of time, check if the path is clear, and continue moving (Geiger, 2022). While a fair criticism to such a step-by-step approach would be its rigidity, in safety-critical scenarios, we cannot afford random behavior. Still, there's room for concern that such an approach does not scale, i.e. there are too many situations for which the machine's behavior must be specified, however, the complexity here is determined by the level of granularity (van Steenkiste et al., 2019) at which step-by-step reasoning must be dictated. In any case, this approach to high-level reasoning is symbolic processing, and thereby assumes the perception module outputs state variables, or *symbols*.

While we have discussed at length the annotation bottleneck for validating assumptions and evaluating disentanglement, in particular what we can do with already annotated datasets (Idrissi et al., 2023) and how we can bridge the gap between proxy metrics (Everingham et al., 2015) and disentanglement, it remains the case that the goal of disentanglement is self-supervised (Balestriero et al., 2023) recovery of the FoVs (Bengio et al., 2013). In that sense, disentanglement (Higgins

et al., 2018) is critical for resolving the failures of symbolic AI (Marcus, 2020), by giving us the symbols we desire without requiring the impractical cost of annotating such a dataset ourselves, building a photorealistic simulation (Bordes et al., 2023b) that closes the sim2real gap well enough that it yields model performance that is sufficiently reliable, or generating synthetic data (Wiles et al., 2022) from unreliable models (Zhang et al., 2024b; Wang et al., 2024). We are excited not only by the progress, but also the prospect of what disentanglement promises.

3.3 Why are representations linear?

What is the state of progress? In (Von Kügelgen et al., 2021), we saw that state-of-the-art approaches to visual SSL do not yield disentangled representations, but instead representations contain all and only information about *content*, i.e., the FoVs invariant to the data augmentations applied in practice. Specifically, there exists an invertible function between the representation and the content variables. We empirically confirmed this result for contrastive (Chen et al., 2020) and non-contrastive (Zbontar et al., 2021) representation learning. Note that this approach of aligning data augmentations of the same image still lies at the heart of discriminative self-supervised pre-training (Oquab et al., 2023; Caron et al., 2021), though, as discussed, the use of multi-crop (Caron et al., 2020) may lead to different results. In any case, our results have proven to be generally applicable; (Daunhawer et al., 2023) and (Liu et al., 2024) applied the results of (Von Kügelgen et al., 2021) and (Zimmermann et al., 2021), respectively, to multimodal contrastive learning (Radford et al., 2021). (Yao et al., 2024) further extended the results of (Von Kügelgen et al., 2021) to an arbitrary number of views and modalities. Finally, (Xu et al., 2024) further extended the results of (Yao et al., 2024) by showing promising results for disentanglement (Zimmermann et al., 2021; Klindt et al., 2021) in the multi-environment setting.

Despite their use in the recent literature, it should be noted that, in terms of representation learning in practice, the results only apply to SSL methods that maximize alignment (Wang and Isola, 2020) while avoiding collapsed representations through regularization (Chen and He, 2021). SSL (Balestriero et al., 2023), however, contains many practical methods beyond that scope, most notably, generative models (Bengio et al., 2000; Brown et al., 2020; Ramesh et al., 2021), which model the data distribution for data sampling (Song and Ermon, 2019). The latest models are multimodal (Radford et al., 2021; Ramesh et al., 2021; Rombach et al., 2022), especially since conditioning generative models on textual input is useful for controllability. While we have considered generative approaches in our theoretical and empirical work (Klindt et al., 2021; Von Kügelgen et al., 2021), given our interest in representation learning, in practice, the generative approaches we've studied are VAEs (Kingma and Welling, 2013). While VAEs have been connected to modern generative models (Luo, 2022; Dieleman, 2022; Chen et al., 2024), diffusion models (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020) in particular, we cannot claim that our results explain what modern generative models represent.

However, given their societal relevance (Eloundou et al., 2023; Cooper et al., 2023; Barrett et al., 2023), a significant amount of recent work has looked into the interpretability of foundation models (Bommasani et al., 2021), ultimately aiming to not only understanding what representations represent (Von Kügelgen et al., 2021), but also what *mechanisms* led to this representation (Olah, 2022). Interestingly, in the mechanistic interpretability literature (Nanda et al., 2023; Gurnee and Tegmark, 2024), it was found that a linear map can be sufficient for decoding concepts from language model representations, e.g. that the output language is French, implying that, at least in the aforementioned experiments, the relationship between extracted representations and the information of interest is linear. (Park et al., 2023) drew an equivalence between the existence of a linear probe and two other phenomena. First, the interesting findings

of vector arithmetic for word embeddings (Mikolov et al., 2013; Pennington et al., 2014); if, in representation space, "woman"-"man", "king"-"queen", and so on are all parallel vectors, then that direction in vector space can be used as a linear probe for Male/Female. Second, the recent use of steering vectors, demonstrated for both diffusion models (Wang et al., 2023) and language models (Turner et al., 2023); if, in representation space, "He is the monarch of England"-"She is the monarch of England", and so on is consistent for Male/Female, then adding said vector to the representation of the context changes the probability of Male/Female. Finally, (Park et al., 2023) showed empirical evidence for linear representations in LLaMA-2-7B (Touvron et al., 2023), a decoder-only Transformer language model (Vaswani et al., 2017; Radford et al., 2018). Theoretical support for linear representations in foundation models has followed, with (Jiang et al., 2024) attributing this phenomenon to next token prediction and the implicit bias of gradient descent, while (Rajendran et al., 2024) shows that data diversity can enable provable recovery of the concepts.

What open problems remain? In the identifiability literature (Khemakhem et al., 2020a,b), weak identifiability has referred to identification up to a linear transformation, while strong identifiability has referred to identification up to a scaled permutation. Given we're interested in disentanglement, our work has aimed to invert the data generating process (Zimmermann et al., 2021) in order to identify the FoVs (Bengio et al., 2013), and thus we have treated strong identifiability as our criterion, since, with weak identifiability, the FoVs are still mixed. Considering the recent evidence that state-of-the-art models possess "linear representations" (Park et al., 2023; Jiang and Aragam, 2023; Rajendran et al., 2024), the results showed that, for a given binary concept, e.g. Male/Female, there exists a linear map for decoding said concept from the language model representation. Note that this is a weaker result than weak identifiability, as weak identifiability implies that there is a single linear map for decoding all FoVs from the representation. However, given all FoVs are unknown, the notion of "linear representations" (Park et al., 2023; Jiang and Aragam, 2023; Rajendran et al., 2024) can be seen as a practical instantiation of weak identifiability for studying real-world models. This begs the question, why the discrepancy in results? Why do we observe a practical form of weak identifiability for language models (Park et al., 2023), when we've seen evidence that state-of-the-art approaches to visual SSL discard FoVs variant to the data augmentations applied in practice (Von Kügelgen et al., 2021)?

First, if we consider our theoretical result for generative SSL in (Von Kügelgen et al., 2021), we did not find that FoVs variant to augmentations, i.e. style, were discarded from the representation, but instead that that the block of style variables was simply separated, or disentangled, from the block of content variables. The discarding of style only occurred for discriminative SSL with non-invertible encoders. Second, we found that the discarding of style is dependent on the relationship between the dimensionality of the encoder output and the number of content FoVs; as the encoder output dimensionality exceeds this quantity, we can observe a clear increase in the amount of style information encoded. Finally, our result only applies to the encoder output, when, in practice, the encoder output is treated as a projection of an intermediate layer, and it's the intermediate layer that's used downstream, as the intermediate layer preserves more information across FoVs (Chen et al., 2020; Von Kügelgen et al., 2021). We found that our ability to linearly decode style variables significantly increased if we evaluated an intermediate layer instead of the output layer, e.g. for data augmentations used in practice (Chen et al., 2020), the R^2 coefficient of determination for linearly decoding object position increased from 0.35 for the output layer to 0.71 for an intermediate layer. Altogether, despite the fact that maximizing alignment with data augmentations does clearly bias visual SSL to discard style, we do see evidence that suggests that style may still be linearly decodable from model representations, if only approximately.

Moreover, note that state-of-the-art approaches to visual SSL are not restricted to approaches

that maximize alignment for isolating content. While DINOv2 (Oquab et al., 2023) primarily relies on the DINO (Caron et al., 2021) objective for performance, which does fit the mold of methods covered by our results for discriminative SSL with non-invertible encoders (Von Kügelgen et al., 2021), as discussed, the use of multi-crop (Caron et al., 2020) may rule out the shortcut (Geirhos et al., 2020) solution due to forcing all FoVs to be variant, or change. Furthermore, while CLIP (Radford et al., 2021; Cherti et al., 2023) has also been described by an application of our results (Daunhawer et al., 2023; Von Kügelgen et al., 2021), in the multimodal setting, all concepts of interest may be shared across modalities, often what's deemed important about an image is written in the caption, and thus learning to discard style may no longer be of concern for linearly decoding concepts of interest, as concepts of interest could simply be content. Lastly, the JEPA models (LeCun, 2022; Baevski et al., 2022; Assran et al., 2023) are competitive with the state-of-the-art (Bardes et al., 2023), and perform masked feature prediction (Vincent et al., 2008; Devlin et al., 2018; He et al., 2022). Masked prediction is inherently a generative modeling task, given the objective is to generate missing parts conditioned on the observed ones, which isn't fundamentally different from language modeling (Balestriero et al., 2023). Thus, while "linear representations" (Park et al., 2023) haven't yet been shown for state-of-the-art approaches to representation learning in visual SSL, it is plausible that the results will extend to this domain.

What is the way forward? While "linear representations" (Park et al., 2023) do not output symbols for the FoVs directly (Marcus, 2020), and thus do not solve the problem of disentangled representation learning, the fact that they are observed in practice from the most highly capable models is promising for the route of translating mechanistic understanding (Olah, 2022) into resolutions to the long-standing issues that deep learning faces (Marcus, 2020).

Let us consider the work of (Wang et al., 2023) as a case study. Here, the problem is compositional generation (Leivada et al., 2023; Okawa et al., 2023; Du et al., 2021), i.e. the tendency for the model to fail unpredictably when presented with compositional inputs, e.g. "magenta-colored panda". While recent work (Rassin et al., 2023) has introduced test-time repairs for many of these binding failure cases, these repairs are problem-specific, e.g. specifically address problems with binding entity-nouns to their modifiers. A preferable solution would be to leverage an improved understanding of the network's processing to offer a more generally applicable repair. (Wang et al., 2023), inspired by vector arithmetic (Mikolov et al., 2013), acts on the "linear representation hypothesis" (Park et al., 2023) by steering text-to-image diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022) to generate counterfactuals (Locatello et al., 2020a; Von Kügelgen et al., 2021; Ahuja et al., 2022a; Sauer and Geiger, 2021) by leveraging the projection onto the subspace that has been identified to correspond to the concept being intervened on. In doing so, the representation of the input text prompt is unchanged except on the targeted subspace, where the representation takes on the value elicited by the intervention (Wang et al., 2023). The score function (Song and Ermon, 2019) of diffusion models (Ho et al., 2020) is used as the representation. This approach was able to outperform Composable-Diffusion (Liu et al., 2022) due to the insight that the score representations of concepts should only be combined after projecting onto the target subspace, an insight derived from concepts corresponding to subspaces of the representation space (Mikolov et al., 2013; Pennington et al., 2014), which implies "linear representation".

Looking forward, identifying linear maps for decoding binary concepts from language model representations is still far from achieving the goals of mechanistic interpretability (Olah, 2022), as how and why the mechanisms prior to the linear representation yielded a linear representation remains unclear. While ongoing research continues to work towards cracking the black box that is deep learning, it is worth noting the value that theory provides. Namely, when a theorem is proven, the assumptions imply the conditions that must be satisfied for the successful result to

occur, thus not only indicating when this result occurs, but also providing the set of conditions that should be investigated if a more satisfactory explanation for the examined phenomenon is desired. An empirical approach to theoretical understanding (Nakkiran, 2021) necessitates thorough, systematic experimentation to understand what experimental parameters are causal for the phenomenon of interest. If mechanistic interpretability research proceeds in that direction, case studies such as (Wang et al., 2023) suggest that investigations into directions such as the "linear representation hypothesis" (Park et al., 2023) could lead to an understanding of deep learning that enables engineering trustworthy, reliable learning systems.

Bibliography

Safe, secure, and trustworthy development and use of artificial intelligence. `https://www.federalregister.gov/d/2023-24283`, 2023. Executive Order 14110. Cited on page 1.

Arash Afkanpour, Vahid Reza Khazaie, Sana Ayromlou, and Fereshteh Forghani. Can generative models improve self-supervised representation learning? *arXiv preprint arXiv:2403.05966*, 2024. Cited on page 117.

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022a. Cited on pages 115, 117, and 123.

Kartik Ahuja, Divyat Mahajan, Vasilis Syrgkanis, and Ioannis Mitliagkas. Towards efficient representation identification in supervised learning. In *Conference on Causal Learning and Reasoning*, pages 19–43. PMLR, 2022b. Cited on page 115.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023a. Cited on pages 115 and 117.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. *arXiv preprint arXiv:2310.02854*, 2023b. Cited on page 115.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. Cited on page 123.

Görkay Aydemir, Weidi Xie, and Fatma Guney. Self-supervised object-centric learning for videos. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 119.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. Cited on page 117.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. Cited on page 123.

Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, et al. A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210*, 2023. Cited on pages 1, 116, 120, 121, and 123.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. Cited on page 2.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. Cited on page 123.

Clark Barrett, Brad Boyd, Ellie Burzstein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *arXiv preprint arXiv:2308.14840*, 2023. Cited on page 121.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000. Cited on page 121.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. Cited on pages 1, 3, 120, and 122.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 118.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. *arXiv preprint arXiv:2311.02695*, 2023. Cited on page 115.

Ondrej Biza, Sjoerd Van Steenkiste, Mehdi SM Sajjadi, Gamaleldin Fathy Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In *International Conference on Machine Learning*, pages 2507–2527. PMLR, 2023. Cited on page 119.

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2021. Cited on page 118.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. Cited on page 121.

Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *Transactions on Machine Learning Research*, 2022. Cited on page 118.

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*, 2023a. Cited on page 119.

Florian Bordes, Shashank Shekhar, Mark Ibrahim, Diane Bouchacourt, Pascal Vincent, and Ari Morcos. Pug: Photorealistic and semantically controllable synthetic data for representation learning. *Advances in Neural Information Processing Systems*, 36, 2023b. Cited on pages 117 and 121.

Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=QO9-y8also-`. Cited on page 1.

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius Von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning*, pages 3038–3062. PMLR, 2023. Cited on pages 116, 117, and 119.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022. Cited on pages 115, 117, and 118.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. Cited on pages 1 and 121.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018. Cited on page 1.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. *Advances in Neural Information Processing Systems*, 35:16946–16961, 2022. Cited on page 116.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 115.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. Cited on pages 116 and 119.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. Cited on pages 119, 121, and 123.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. Cited on pages 2, 119, 120, 121, and 123.

Michael Chang, Tom Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. *Advances in Neural Information Processing Systems*, 35:32694–32708, 2022. Cited on page 119.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. Cited on pages 1, 2, 118, 119, 120, 121, and 122.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. Cited on pages 2 and 121.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024. Cited on page 121.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. Cited on page 123.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. Cited on pages 1 and 3.

A Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, et al. Report of the 1st workshop on generative ai and law. *arXiv preprint arXiv:2311.06477*, 2023. Cited on page 121.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021. Cited on page 117.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. Cited on page 119.

George Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle linéaire. *Revue de l'Institut international de statistique*, pages 2–8, 1953. Cited on pages 115 and 116.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on pages 118, 121, and 123.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. Cited on page 118.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. Cited on pages 119 and 123.

Sander Dieleman. Diffusion models are autoencoders, 2022. URL `https://benanne.github.io/2022/01/31/diffusion.html`. Cited on page 121.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wuthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. In *International Conference on Learning Representations*, 2021. Cited on page 118.

Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. Cited on page 2.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`. Cited on page 119.

Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34:15608–15620, 2021. Cited on page 123.

Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Bernhard Schölkopf, and Mark Ibrahim. Self-supervised disentanglement by leveraging structure in data augmentations. *arXiv preprint arXiv:2311.08815*, 2023. Cited on pages 3, 117, and 119.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023. Cited on pages 1 and 121.

Gamaleldin Elsayed, Aravindh Mahendran, Sjoerd Van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. *Advances in Neural Information Processing Systems*, 35:28940–28954, 2022. Cited on page 119.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. Cited on page 1.

Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. Cited on pages 1 and 120.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023. Cited on page 117.

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. Cited on page 116.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 118.

Christina M Funke, Paul Vicol, Kuan-Chieh Wang, Matthias Kümmerer, Richard Zemel, and Matthias Bethge. Disentanglement and generalization under correlation shifts. In *Conference on Lifelong Learning Agents*, pages 116–141. PMLR, 2022. Cited on page 117.

Andreas Geiger. Self-Driving Cars Lecture Notes. PDF document, 2022. URL `https://drive.google.com/file/d/1N7aPV1xHVYqm25OfsfSIVusZ6ERkYmas/view`. Cited on page 120.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. Cited on pages 4, 5, 116, and 118.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`. Cited on page 1.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. Cited on pages 119 and 123.

Robert Geirhos, Roland S Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719*, 2023. Cited on page 1.

Shubhangi Ghosh, Luigi Gresele, Julius von Kügelgen, Michel Besserve, and Bernhard Schölkopf. Independent mechanism analysis and the manifold hypothesis. *arXiv preprint arXiv:2312.13438*, 2023. Cited on page 116.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32, 2019. Cited on page 118.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. Cited on page 1.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International conference on machine learning*, pages 2424–2433. PMLR, 2019. Cited on pages 118 and 119.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. Cited on page 116.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. Cited on page 118.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in neural information processing systems*, 34: 28233–28248, 2021. Cited on page 116.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. Cited on page 2.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 121.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. Cited on page 2.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. Cited on page 119.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. Cited on page 123.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017a. URL `https://openreview.net/forum?id=Sy2fzU9gl`. Cited on pages 117 and 118.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, pages 1480–1490. PMLR, 2017b. Cited on page 118.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. Cited on pages 116, 119, and 120.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. Cited on pages 121 and 123.

Kyle Hsu, William Dorrell, James Whittington, Jiajun Wu, and Chelsea Finn. Disentanglement via latent quantization. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 117.

Xinyu Huang, Yi-Jie Huang, Youcai Zhang, Weiwei Tian, Rui Feng, Yuejie Zhang, Yanchun Xie, Yaqian Li, and Lei Zhang. Inject semantic concepts into image tagging for open-set recognition. *arXiv preprint arXiv:2310.15200*, 2023. Cited on page 117.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985. Cited on page 118.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016. Cited on pages 115 and 118.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017. Cited on pages 115 and 118.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. Cited on pages 1, 115, and 116.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019. Cited on pages 115, 117, and 118.

Aapo Hyvärinen, Ilyes Khemakhem, and Hiroshi Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *Patterns*, 4(10), 2023. Cited on page 1.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on pages 117, 118, and 120.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020. Cited on page 120.

Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010. Cited on page 116.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on pages 115 and 122.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. *arXiv preprint arXiv:2403.03867*, 2024. Cited on page 122.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. Cited on page 119.

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. *Advances in Neural Information Processing Systems*, 34:20146–20159, 2021. Cited on page 118.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020a. Cited on pages 115, 118, and 122.

Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020b. Cited on page 122.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. Cited on page 117.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. Cited on pages 116 and 121.

Thomas Kipf, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022. Cited on page 118.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. Cited on page 117.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022. Cited on page 117.

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=EbIDjBynYJ8`. Cited on pages 4, 5, 115, 116, 118, 119, 120, and 121.

David Alexander Klindt, Aapo Hyvarinen, Axel Levy, Nina Miolane, and Frederic Poitevin. Towards interpretable cryo-em: Disentangling latent spaces of molecular conformations. *bioRxiv*, pages 2024–03, 2024. Cited on page 118.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. Cited on page 115.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. Cited on page 1.

Abhishek Kumar and Ben Poole. On implicit regularization in beta-vaes. In *International Conference on Machine Learning*, pages 5480–5490. PMLR, 2020. Cited on page 116.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022. Cited on page 115.

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. *Advances in Neural Information Processing Systems*, 36, 2023a. Cited on pages 116 and 119.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024. Cited on page 115.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *The Eleventh International Conference on Learning Representations*, 2023b. Cited on page 118.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022. Cited on pages 1 and 123.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. Cited on page 1.

Evelina Leivada, Elliot Murphy, and Gary Marcus. Dall · e 2 fails to reliably capture common syntactic processes. *Social Sciences & Humanities Open*, 8(1):100648, 2023. Cited on page 123.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. Cited on page 119.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022. Cited on pages 115 and 118.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. In *Uncertainty in Artificial Intelligence*, pages 1263–1273. PMLR, 2023a. Cited on page 115.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023b. Cited on page 115.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022. Cited on page 123.

Nan Liu, Yilun Du, Shuang Li, Joshua B Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2095, 2023a. Cited on page 118.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning*, pages 553–573. PMLR, 2023b. Cited on page 118.

Yuejiang Liu, Ahmad Rahimi, Po-Chien Luan, Frano Rajič, and Alexandre Alahi. Sim-to-real causal transfer: A metric learning approach to causally-aware interaction representations. *arXiv preprint arXiv:2312.04540*, 2023c. Cited on page 118.

Yuhang Liu, Zhen Zhang, Dong Gong, Biwei Huang, Mingming Gong, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Revealing multimodal contrastive representation learning through latent partial causal models. *arXiv preprint arXiv:2402.06223*, 2024. Cited on page 121.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. Cited on pages 115, 116, and 117.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR, 2020a. Cited on pages 115 and 123.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020b. Cited on pages 116 and 119.

Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023. Cited on page 118.

Sindy Löwe, Phillip Lippe, Francesco Locatello, and Max Welling. Rotating features for object discovery. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 119.

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. Cited on page 121.

Gary Marcus. The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020. Cited on pages 121 and 123.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. Cited on page 118.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. Cited on pages 122 and 123.

Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. MOT16: A benchmark for multi-object tracking. *arXiv:1603.00831 [cs]*, March 2016. Cited on pages 4, 5, 116, and 118.

Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020. Cited on page 119.

Milton Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, 35:10136–10149, 2022. Cited on page 116.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. Cited on pages 116 and 117.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Deepdream-a code example for visualizing neural networks. *Google Research*, 2(5), 2015. Cited on page 1.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *arXiv preprint arXiv:2310.15709*, 2023. Cited on page 115.

Preetum Nakkiran. *Towards an empirical theory of deep learning*. PhD thesis, Harvard University, 2021. Cited on page 124.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, 2023. Cited on page 121.

Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 123.

Chris Olah. Mechanistic interpretability, variables, and the importance of interpretable bases, 2022. URL `https://transformer-circuits.pub/2022/mech-interp-essay/`. Cited on pages 121 and 123.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. Cited on page 1.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. Cited on page 2.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. Cited on pages 2, 121, and 123.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. Cited on page 1.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023. Cited on pages 121, 122, 123, and 124.

Nikhil Parthasarathy, SM Ali Eslami, Joao Carreira, and Olivier J Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. Cited on page 5.

Judea Pearl. *Causality*. Cambridge university press, 2009. Cited on pages 115 and 117.

William Peebles, John Peebles, Jun-Yan Zhu, Alexei Efros, and Antonio Torralba. The hessian penalty: A weak prior for unsupervised disentanglement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 581–597. Springer, 2020. Cited on page 116.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. Cited on pages 122 and 123.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. Cited on pages 115 and 116.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. Cited on page 120.

Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics meets temporal correspondence: Self-supervised object-centric learning in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16675–16687, 2023. Cited on page 119.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. Cited on page 122.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. Cited on pages 117, 121, and 123.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning interpretable concepts: Unifying causal representation learning and foundation models. *arXiv preprint arXiv:2402.09236*, 2024. Cited on page 122.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. Cited on page 121.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 123.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: Vaes perform independent mechanism analysis. *Advances in Neural Information Processing Systems*, 35:12040–12057, 2022. Cited on page 116.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021. Cited on page 119.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12406–12415, 2019. Cited on page 116.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. Cited on pages 121 and 123.

Karsten Roth, Mark Ibrahim, Zeynep Akata, Pascal Vincent, and Diane Bouchacourt. Disentanglement of correlated factors via hausdorff factorized support. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on page 117.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. Cited on page 117.

Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in Neural Information Processing Systems*, 35:9512–9524, 2022. Cited on page 119.

Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. Cited on pages 116 and 123.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 459–466, 2012. Cited on page 116.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021. Cited on pages 115 and 118.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. In *The Eleventh International Conference on Learning Representations*, 2023. Cited on page 119.

Yash Sharma, Yi Zhu, Chris Russell, and Thomas Brox. Pixel-level correspondence for self-supervised learning from video. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022. Cited on page 5.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *International Conference on Learning Representations*, 2022a. Cited on page 119.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in Neural Information Processing Systems*, 35:18181–18196, 2022b. Cited on page 119.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020. Cited on page 1.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. Cited on pages 121 and 123.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. Cited on pages 121 and 123.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pages 32540–32560. PMLR, 2023. Cited on pages 115 and 117.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 115.

Mikhail Fedorovich Subbotin. On the law of frequency of error. *Mat. Sb.*, 31(2):296–301, 1923. Cited on page 4.

Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European conference on computer vision*, pages 438–451. Springer, 2010. Cited on page 120.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. Cited on page 119.

Davide Talon, Phillip Lippe, Stuart James, Alessio Del Bue, and Sara Magliacane. Towards the reusability and compositionality of causal representations. *arXiv preprint arXiv:2403.09830*, 2024. Cited on page 118.

Matthias Tangemann, Steffen Schneider, Julius Von Kügelgen, Francesco Locatello, Peter Vincent Gehler, Thomas Brox, Matthias Kuemmerer, Matthias Bethge, and Bernhard Schölkopf. Unsupervised object learning via common fate. In *Conference on Causal Learning and Reasoning*, pages 281–327. PMLR, 2023. Cited on page 120.

Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023a. Cited on page 117.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2023b. Cited on page 117.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. Cited on page 122.

Michael Tschannen, Josip Djolonga, Marvin Ritter, Aravindh Mahendran, Neil Houlsby, Sylvain Gelly, and Mario Lucic. Self-supervised learning of video-induced visual invariances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13806–13815, 2020. Cited on page 5.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023. Cited on page 122.

Sjoerd van Steenkiste, Klaus Greff, and Jürgen Schmidhuber. A perspective on objects and systematic generalization in model-based rl. *arXiv preprint arXiv:1906.01035*, 2019. Cited on page 120.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *arXiv preprint arXiv:2402.00849*, 2024. Cited on page 115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. Cited on pages 119 and 122.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. Cited on page 123.

Bill Vlasic and Neil E. Boudette. Self-driving tesla was involved in fatal crash, u.s. says. `https://www.nytimes.com/2016/07/01/business/self-driving-tesla-fatal-crash-investigation.html`, 2016. Cited on page 1.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. Cited on pages 4, 5, 116, and 118.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. Cited on pages 2, 5, 115, 117, 118, 119, 121, 122, and 123.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on pages 115 and 117.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. Cited on page 118.

Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. Do clips always generalize better than imagenet models? *arXiv preprint arXiv:2403.11497*, 2024. Cited on page 121.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. Cited on pages 2, 3, and 121.

Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021. Cited on page 117.

Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on pages 122, 123, and 124.

Nicholas Watters, Loic Matthey, Sebastian Borgeaud, Rishabh Kabra, and Alexander Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment, 2019. URL `https://github.com/deepmind/spriteworld/`. Cited on page 118.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 116.

Olivia Wiles, Isabela Albuquerque, and Sven Gowal. Discovering bugs in vision models using off-the-shelf image generation and captioning. *arXiv preprint arXiv:2208.08831*, 2022. Cited on pages 117 and 121.

Markus Wulfmeier, Arunkumar Byravan, Tim Hertweck, Irina Higgins, Ankush Gupta, Tejas Kulkarni, Malcolm Reynolds, Denis Teplyashin, Roland Hafner, Thomas Lampe, et al. Representation matters: Improving perception and exploration for robotics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6512–6519. IEEE, 2021. Cited on pages 116, 119, and 120.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024. Cited on pages 118 and 121.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. Cited on pages 4, 5, 116, and 118.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *arXiv preprint arXiv:1905.04804*, 2019. Cited on pages 4, 5, 116, 118, 119, and 120.

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 121.

Andrii Zadaianchuk, Matthaeus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *The Eleventh International Conference on Learning Representations*, 2023a. Cited on page 119.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. *Advances in Neural Information Processing Systems*, 36, 2023b. Cited on pages 119 and 120.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. Cited on pages 2, 3, and 121.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. Cited on page 1.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on page 115.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024a. Cited on page 115.

Yizhe Zhang, He Bai, Ruixiang Zhang, Jiatao Gu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. How far are we from intelligent visual deductive reasoning? *arXiv preprint arXiv:2403.04732*, 2024b. Cited on page 121.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in Neural Information Processing Systems*, 35:16411–16422, 2022. Cited on pages 116 and 117.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. Cited on pages 3, 5, 115, 116, 118, 119, 120, 121, and 122.