How motion promotes perceptual organization in humans and machines

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

vorgelegt
von
Matthias Tangemann
aus Stuttgart

Tübingen
2025

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

Tag der mündlichen Qualifikation:	24.07.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Matthias Bethge
2. Berichterstatter:	Prof. Dr. Jakob Macke

Ich erkläre hiermit, dass ich die zur Promotion eingereichte Arbeit mit dem Titel *The role of motion in perceptual organization: bridging human and machine vision* selbständig verfasst, nur die angegebenen Quellen und Hilfsmittel benutzt und wörtlich oder inhaltlich übernommene Stellen als solche gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.5.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Tübingen, den _____ _____

Datum / Date Unterschrift / Signature

Acknowledgments

This thesis would not have been possible without many people that accompanied and supported my work during the last years in many ways.

First and foremost, I would like to thank my supervisor Matthias Bethge for his inspiration, motivation, and trust in my work. Research related to machine learning is changing quickly and often overwhelming. I consider myself extremely lucky to have found an environment in which I had the freedom to explore and develop my own scientific standpoint, while simultaneously having the opportunity and encouragement to engage with the leading developments in our field. Thank you, Matthias, for creating this inspiring and supportive environment.

Throughout my entire Ph.D., Matthias Kümmerer provided guidance and support as a postdoctoral researcher. I am grateful for the opportunity to learn how to navigate day-to-day academic life as well as many productive work and non-work related discussions. Thank you for your support and friendship throughout these years.

I would like to thank Felix Wichmann for many opportunities to learn about human vision and encouraging feedback on my work. Your perspective uniquely contributed to my scientific development and to shaping my future ambitions.

Furthermore, I would like to thank all my coauthors for contributing their knowledge and time to the various projects. Moreover, I am thankful to all members of the Bethgelab and the adjacent scientific community in Tübingen for the great time we spent together in the lab, after work and while traveling. I am no less thankful for the constant support from the administrative team, and in particular Heike König, for helping with all organizational aspects of working at the university.

Last but not least, I am deeply grateful for my family. Thank you for always supporting me, providing a safe space of normality during stressful times and for reminding me of the good life beyond science.

Summary

Perceptual organization is a core process of human vision that transforms the raw visual input into a structured, object-centric scene representation. Motion information plays a central role in this process: On the one hand, perceiving motion requires a perceptual organization process to establish an identity behind percepts at different time points. On the other hand, motion information has been shown to be a dominant cue that humans use to infer scene structure, for example by the Gestalt principle of common fate. Moreover, motion has been shown to enable more efficient processing by guiding attention to relevant areas of scenes, and to contribute to learning perceptual organization during infancy.

In this thesis, we combine insights from psychology and neuroscience with recent advances in machine learning in order study how motion motion promotes different aspects of dynamic scene perception. First, we study the role of motion in guiding eye movements as a basis for more efficient scene perception. Our analysis reveals several strong effects of temporal patterns on eye movements in a data-driven manner, but also identifies their scarcity in common benchmarks as a key limitation for modeling this process. We propose a new benchmark that combines the respective cases from several existing benchmarks to support future research on this topic. In our second project, we take inspiration from developmental psychology and study the role of motion for learning how to decompose a scene into objects. Trained this way, our model reflects central capabilities of scene perception in humans, such as the ability to complete partial objects and to generate novel scenes that systematically generalize beyond the training distribution. Finally, we study the neural basis of motion segmentation using a combination of computational modeling and experimental psychophysics. We find striking differences between state-of-the-art computer vision models and human perception in terms of appearance-independent segmentation of moving random dot patterns. Furthermore, we show that a neuroscience-inspired motion energy approach allows matching human perception and thus provide a compelling link between the neural mechanisms of motion perception and the Gestalt principle of common fate.

In summary, the projects in this thesis contribute to our understanding how motion information promotes perceptual organization from an interdisciplinary NeuroAI perspective. DNNs allow building more capable scientific models of human vision, and thus enable novel insights into the perception of natural scenes. Conversely, we show that insights from human vision can be successfully transferred to a computer vision setting. Our work therefore contributes to a more holistic understanding of human vision and provides insights that may inspire more capable machine vision in the future.

Zusammenfassung

Die Wahrnehmungsorganisation ist ein Kernprozess des menschlichen Sehens, der den rohen visuellen Input in eine strukturierte, objektzentrierte Szenendarstellung transformiert. Bewegungsinformationen spielen in diesem Prozess eine zentrale Rolle: Einerseits erfordert die Wahrnehmung von Bewegung einen Wahrnehmungsorganisationsprozess, um eine Identität hinter Wahrnehmungen zu verschiedenen Zeitpunkten herzustellen. Andererseits hat sich gezeigt, dass Bewegungsinformationen ein dominanter Hinweis sind, den Menschen nutzen, um die Struktur einer Szene zu erschließen, beispielsweise durch das Gestaltprinzip des gemeinsamen Schicksals. Darüber hinaus hat sich gezeigt, dass Bewegung effizientere Verarbeitung ermöglicht, indem sie die Aufmerksamkeit auf relevante Bereiche von Szenen lenkt und zur Erlernung der Wahrnehmungsorganisation in der frühen Kindheit beiträgt.

In dieser Dissertation kombinieren wir Erkenntnisse aus der Psychologie und Neurowissenschaft mit den jüngsten Fortschritten im maschinellen Lernen, um zu untersuchen, wie Bewegung verschiedene Aspekte der dynamischen Szenenwahrnehmung fördert. Zuerst untersuchen wir die Rolle von Bewegung bei der Steuerung von Augenbewegungen als Grundlage für eine effizientere Szenenwahrnehmung. Unsere Analyse zeigt mehrere starke Effekte von zeitlichen Mustern auf Augenbewegungen auf datengetriebene Weise auf, identifiziert aber auch deren Knappheit in gängigen Benchmarks als eine wesentliche Einschränkung für die Modellierung dieses Prozesses. Wir schlagen einen neuen Benchmark vor, der die jeweiligen Fälle aus mehreren bestehenden Benchmarks kombiniert, um die zukünftige Forschung zu diesem Thema zu unterstützen. In unserem zweiten Projekt lassen wir uns von der Entwicklungspsychologie inspirieren und untersuchen die Rolle der Bewegung beim Erlernen der Zerlegung einer Szene in Objekte. Auf diese Weise trainiert, reflektiert unser Modell zentrale Fähigkeiten der Szenenwahrnehmung bei Menschen, wie die Fähigkeit, unvollständige Objekte zu vervollständigen und neue Szenen zu generieren, die systematisch über die Trainingsverteilung hinaus generalisieren. Schließlich untersuchen wir die neuronale Grundlage der Bewegungssegmentierung mittels einer Kombination aus computergestütztem Modellieren und experimenteller Psychophysik. Wir finden auffällige Unterschiede zwischen modernen Computervisionsmodellen und der menschlichen Wahrnehmung in Bezug auf das erscheinungsunabhängige Segmentieren von sich bewegenden Zufallspunktmustern. Darüber hinaus zeigen wir, dass ein neurowissenschaftlich inspirierter Bewegungsenergieansatz die menschliche Wahrnehmung nachbilden kann und somit eine überzeugende Verbindung zwischen den neuronalen Mechanismen der Bewegungswahrnehmung und dem Gestaltprinzip des gemeinsamen Schicksals bietet.

Zusammenfassend tragen die Projekte in dieser Dissertation zu unserem Verständnis bei, wie Bewegungsinformationen die Wahrnehmungsorganisation aus einer interdisziplinären NeuroAI-Perspektive fördern. DNNs ermöglichen den Aufbau fähigerer wissenschaftlicher Modelle des menschlichen Sehens und bieten somit neue Einblicke in die Wahrnehmung natürlicher Szenen. Umgekehrt zeigen wir, dass Erkenntnisse aus dem menschlichen Sehen erfolgreich in einen Computervisionskontext übertragen werden können. Unsere Arbeit trägt daher zu einem ganzheitlicheren Verständnis des menschlichen Sehens bei und bietet Erkenntnisse, die möglicherweise fähigere maschinelle Sicht in der Zukunft inspirieren könnten.

The summary was automatically translated to German using ChatGPT (https://chat.openai.com) and proofread by the author

Contents

List of publications

The following peer-reviewed publications are included in this thesis:

- Matthias Tangemann, Matthias Kümmerer, Thomas S.A. Wallis, Matthias Bethge (2020). *Measuring the importance of temporal features in video saliency.* Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol 12373. Springer, Cham.
- Matthias Tangemann, Steffen Schneider, Julius von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, Bernhard Schölkopf (2023). *Unsupervised object learning via common fate.* Proceedings of the Second Conference on Causal Learning and Reasoning, PMLR 213:281-327.
- Matthias Tangemann, Matthias Kümmerer, Matthias Bethge (2024). *Object segmentation from common fate: Motion energy processing enables human-like zero-shot generalization to random dot stimuli.* Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

More detailed information about the contributions of each author are listed in the chapters discussing the respective publication.

1 Introduction

We perceive the visual world in a highly structured manner. When we look around, we see a hierarchical structure of objects and object parts and the relations between different objects. We see a world that is three dimensional, infer how objects continue behind occluders and estimate their physical properties. Yet, the incoming visual information is very unstructured: Millions of photoreceptor cells in our eyes measure incoming light at every instant and provide a flood of isolated information about the environment. Structuring this low-level visual information into high-level scene percepts requires an active process known as perceptual organization (Wagemans, 2015).

The study of perceptual organization has a long history in psychology, neuroscience and computational modeling. In the early 20th century, Gestalt psychology emerged as an influential conception of perception from the study of motion perception (Wertheimer, 1912) and pioneered the empirical study of perceptual organization (Wagemans, Elder, et al., 2012; Wagemans, Feldman, et al., 2012). The Gestalt psychologists formulated a series of principles that describe how human perception groups elements in a scene and how foreground objects are segregated from background. Supported by ever improving recording techniques, neuroscience has complemented psychology during the second half of the 20th century with a focus on understanding the neural mechanisms that implement human visual perception (Kandel et al., 2021). In parallel, the steady advancements of computers allowed to both contribute to the study of human vision using computational models and increased interest in the underlying mechanisms driven by applications in computer vision (Dayan & Abbott, 2001; Doerig et al., 2023; Hassabis et al., 2017). As such, the study of perceptual organization is inherently multidisciplinary and a full understanding requires insights from all three perspectives.

The perception of motion is deeply connected to perceptual organization in a multi-faceted way. Firstly, perceiving motion requires to establish an identity behind percepts at different time points. Solving this so called *correspondence problem* (Ullman, 1979) is a prime example of a perceptual organization process. Furthermore, motion is a strong cue that humans use for decomposing a scene into objects. The principle of common fate, often paraphrased as "what moves together, belongs together", has been repeatedly shown to strongly influence perceptual organization and motion information has been shown to take precedence over other cues (Z. Huang & Zaidi, 2022; Wertheimer, 1923). Moreover, motion has been shown to be central cue for the development of perceptual organization in infants. Already at a very young age, infants are able to track moving objects and to use motion information to reason about occlusions (Arterberry & Kellman, 2016; Kellman & Spelke, 1983).

It has thus been hypothesized that motion information provides a learning signal for appearance cues of perceptual organization. Finally, motion strongly attracts attention and has been recognized as an essential factor for the efficiency of scene perception by guiding human gaze (Spelke, 1990; Ullman et al., 2012).

While substantial progress in has been made during the last 150 years, a comprehensive understanding of perceptual organization remains elusive. Psychology has demonstrated strong perceptual grouping phenomena using simple stimuli; many questions remain, however, for fully understanding the perception of real world scenes (Võ, 2021). Neuroscience has uncovered mechanisms behind elementary perceptual processes including the perception of motion, but overall we are far from understanding how perceptual organization principles are implemented in the brain (Nishida et al., 2018). Advancing our understanding of how motion promotes perceptual organization of real world scenes is thus an important direction in research and impactful beyond vision science. Humans are visual animals, and the structure of our visual representation is foundational for all higher cognitive functions such as memorization, reasoning and planning. Better understanding perceptual organization is therefore essential for a holistic understanding of the human mind and has direct applications, for example in visual design (Lidwell et al., 2010). Finally, better understanding perceptual organization guides the development of more capable and better aligned computer vision systems (Sundaram et al., 2024).

In this thesis, we present work that adopts an interdisciplinary perspective to examine the role of motion for structured scene perception in humans. The advent of deep learning has enabled a new generation of computer vision models that rival many human visual skills on natural images while showing intriguing parallels to cortical processing in the human brain. As such, deep neural networks are regarded as promising tools to advance our understanding of human vision (e.g., Wichmann and Geirhos, 2023). Most research that relates deep learning and human vision so far has however focused on core object recognition in static images while motion perception and perceptual organization has received much less attention. The projects presented in this thesis contribute to closing this gap and provide novel insights into the mechanisms underlying human perception of dynamic scenes from a computational perspective.

2 Background

2.1 Dynamic scene perception in humans

Motion psychophysics. The study of motion perception is a central topic in psychophysics (Nishida et al., 2018; Wixted & Serences, 2018). It has long been realized that perceiving motion does not require physical motion, as demonstrated by *apparent motion* phenomena in Exner (1875). Rather, motion perception is based on solving the problem of *phenomenological identity* (Ternus, 1926), or *correspondence problem* (Ullman, 1979), that relates perceived patterns across time. During the last century, researchers have systematically studied which stimulus properties are necessary to evoke a percept of motion (e.g., Chubb and Sperling, 1988; Wertheimer, 1912). The variety of conditions under which humans perceive motion has fueled a debate about multiple different processes that interact in motion perception (Braddick, 1974; Mather & Cavanagh, 1989; Petersik, 1991). It seems plausible that two different processes contribute to motion perception: A low-level process that detects motion over the entire visual field, and a higher-level process that is limited by attention to tracking few individual objects (Cavanagh, 1991).

A challenge for any motion perception system is fact that local motion information is ambiguous, known as the *aperture problem*. In the human visual system, elementary motion detection is thus followed by a global integration process that takes into account scene structure in order to resolve these ambiguities (Nishida et al., 2018). This motion interpretation stage has been shown to interact with a diverse set of capabilities such as the segmentation of moving objects (Braddick, 1993), the perception of 3d structure (Ullman, 1997) and the perception of material properties (Doerschner et al., 2011). A particularly striking example for the reach of this higher-level interpretation process is our ability to see biological motion given only few moving dots (Johansson, 1973). Due to the variety of motion phenomena and due to the rich interaction with other aspects of vision, we do not yet have a complete understanding of how humans perceive motion in complex natural scenes (Nishida et al., 2018).

Motion neuroscience. A rich body of work in neuroscience complements research in psychophysics with insights about the neuronal mechanisms that implement motion perception. Direction selective cells, which signal motion in a specific direction, are found across species and in many areas of the visual system (Barlow & Hill, 1963; Mauss et al., 2017). Hassenstein and Reichardt (1956) proposed a *delay-and-compare* mechanism for motion detection based on studying beetles. This so called Reichardt detector has been extended (Barlow & Levick, 1965; van Santen & Sperling, 1985) and experimentally verified in many other animals including

humans (Borst & Egelhaaf, 1989). Another popular line of works approaches motion perception as detection of spatio-temporal orientations using a *motion energy model* (Adelson & Bergen, 1985; Watson & Ahumada, 1985). While this model has a different internal structure, it has been shown to compute the same function as an extended variant of the Reichardt detector (van Santen & Sperling, 1984).

In primates, direction selective cells are found as early as in the retina and V1, but also in higher visual areas such as MT and MST (Albright, 1984; Duffy & Wurtz, 1991). While the cells in V1 respond to elementary motion in a small receptive fields, many cells in MT respond to more complex motion patterns and have much larger receptive fields. It is therefore assumed that these areas act as a motion integration stage and unify the local motion signals into a globally coherent percept. The influential model by Simoncelli and Heeger (1998) extended the motion energy model with a second stage that implements an *intersection-of-constraints* mechanism (Adelson & Movshon, 1982; Fennema & Thompson, 1979) in order to resolve ambiguities. Such feedforward V1-MT models have been shown to explain a range of the firing patterns in visual area MT (Rust et al., 2006; Simoncelli & Heeger, 1998) and have been linked to human perception (Nishimoto et al., 2011; Weiss et al., 2002). Several works have further extended V1-MT models to allow for a more complex integration of local motion beyond spatial averaging (Lidén & Pack, 1999; Wu et al., 2010; Zarei Eskikand et al., 2016).

In summary, the mechanisms of motion detection are among the most extensively studied subjects in neuroscience and a range of computational models have been proposed that explain neural firing (Borst, 2000). While the elementary mechanisms for motion detection are understood very well, we however do not yet have a complete understanding of the mechanisms behind the integration and interpretation of motion patterns in natural scenes (Nishida et al., 2018).

The role of motion for learning perceptual organization. The previous sections have outlined the deep connection between motion perception and perceptual organization in the adult visual system. Additionally, motion information has been shown to be an important cue for learning perceptual organization in infancy (Arterberry & Kellman, 2016). Several studies in developmental psychology revealed distinct error patterns in young infants when perceiving partly occluded objects (Kellman & Spelke, 1983) and an increase of capabilities in terms of visual complexity during the first few months of infancy (Needham, 1998). Strikingly, it has been shown that for moving objects, infant perception agrees with adult perception at a much younger age (Spelke, 1990). It has thus been hypothesized, that infant perception initially structures scenes purely by geometrical cues like motion and depth, and that appearance cues are learned from these geometrical cues during infancy (1990).

While the importance of motion for learning of perceptual organization in infancy has been consistently demonstrated, we're lacking a precise understanding of the learning process that supports computational modeling.

The role of motion for efficient scene perception. Any animal or mobile robot faces the challenge of an excess of visual information that must be processed in real-time under tight energy constraints. In the human eye, only a small central are of the retina, the *fovea*, has a high density of photoreceptors and provides high-resolution visual information (Curcio et al., 1990). Multiple times per second, humans shift their gaze to different locations of the scene and integrate the information into a coherent percept of the surroundings (Yarbus, 1967). Motion plays a central role in this process, which is ecologically plausible since nearby, moving objects are highly relevant for survival. Accordingly, several studies have found a *popout* effect of moving compared to static scene elements (Rosenholtz, 1999; Wolfe, 2000), taking into account self-generated motion by body movements (Rushton et al., 2007). Furthermore, studies on *inattentional blindness* demonstrate that many changes in scenes, such as removed or relocated objects, are not noticed by humans if not attended. Motion plays a central role for the success of this selective processing strategy, since motion signals change in the environment and allows to reallocate attention (Rensink, 2000). A comprehensive understanding of structured scene perception in humans therefore requires understanding how motion influences selective processing through eye movements and covert attention.

2.2 Dynamic scene perception in machines

The field of computer vision has studied how to engineer artificial vision systems since decades (Szeliski, 2022). During recent years, the advent of deep learning has enabled computer vision systems that rival human capabilities for many tasks (LeCun et al., 2015). This recent wave of deep learning was driven by networks trained to classify images of objects (K. He et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2015) but has since been extended and adapted to a diverse set of tasks including the perception of dynamic multi-object scenes. In the following, we outline the most important landmarks as relevant for this thesis.

Object detection and segmentation. The successes of DNNs in image classification directly transferred to object detection by methods that first propose a large number of bounding boxes and then apply a CNN to classify the respective image crops (Girshick et al., 2014). This two-stage approach has been improved over the years (Girshick, 2015; Ren et al., 2015) and was extended to pixel-level segmentation

(K. He et al., 2017; Kirillov et al., 2020). Furthermore, as an alternative to the complex proposal-based methods, single stage architectures have been developed that directly predict bounding boxes (Redmon et al., 2016; Tian et al., 2019) or instance segmentation masks (Carion et al., 2020; Tian et al., 2020). All of these methods rely on pretraining the visual representation for object classification on ImageNet (Deng et al., 2009), followed by supervised finetuning using human-labeled images from common benchmarks such as PASCAL VOC (Everingham et al., 2010) or COCO (Lin et al., 2015). While this supervised approach to object detection and segmentation successfully scales to complex real-world scenes, methods are limited by the amount of human-labeled data and in particular the relatively low number of labeled object classes (Joseph et al., 2021). Prior to the development of the Segment Anything Model SAM (Kirillov et al., 2023) in parallel to this thesis, it was therefore commonly believed that unsupervised approaches were necessary to rival human object segmentation capability.

Object-centric representation learning. As an alternative to supervised instance segmentation, the field of object-centric representation learning aims to develop methods for the compositional representation of multi-object scenes that do not require human labels. Beyond classical instance segmentation, most object-centric methods further attempt to learn a more comprehensive scene representation, including the compositional 3d structure and a structured representation of object attributes and relations between the objects (Greff et al., 2020). Typical approaches are based on autoencoders and trained to reconstruct the input scene (Burgess et al., 2019; Eslami et al., 2016). To foster a decomposition of scenes into objects, these methods rely on architectural biases such as the segregation of the latent representation into *slots* (Locatello et al., 2020) and the interaction of slots using graph neural networks (Battaglia et al., 2018; Kipf et al., 2020). Several works have extended these methods to videos based on a frame-by-frame reconstruction objective, loosely resembling Hidden Markov models (J. Jiang et al., 2020; Kosiorek et al., 2018). Both for images and videos, object-centric representation learning methods have been successfully applied to simple synthetic scenes. However, these models have been shown to fail for more complex scenes and in particular for textured objects (Karazija et al., 2021; Weis et al., 2021), leaving a huge gap to their supervised counterparts.

Optical flow. In computer vision, the estimation of motion is typically framed as estimating the optical flow between two successive frames (Black & Anandan, 1993; Horn & Schunck, 1981). Optical flow estimation has been approached by using CNNs that directly regress motion vectors from the input frames (Dosovitskiy et al., 2015; Ilg et al., 2017) and are trained on large-scale datasets (Mayer et al., 2016).

Further improvements have been enabled by architectures that take inspiration from classical optimization based method (D. Sun et al., 2018; 2020; Teed & Deng, 2020). As an alternative to supervised training on synthetic videos, classical warping-based optimization objectives have been repurposed as a loss for training optical flow estimation networks (Jonschkowski et al., 2020; Meister et al., 2018; Stone et al., 2021). On established benchmarks (Butler et al., 2012; Menze et al., 2015), these methods have reached sub-pixel accuracy and are successfully used for a range of downstream tasks.

Motion segmentation. The recent advancements in segmentation and optical flow estimation have been combined in order to segment moving objects. A classical approach to motion segmentation integrates optical flow into point trajectories spanning multiple frames and performs clustering to find moving segments (Brox & Malik, 2010; Keuper et al., 2015). More recently, deep neural networks for segmentation have been trained with optical flow as input (Lamdouar et al., 2020; 2021; Tokmakov et al., 2017), optionally in combination with RGB input (Dave et al., 2019) or geometrical constraints (Bideau et al., 2018). Due to the strength of image-based segmentation methods and the limited amount of human-labeled videos, many video segmentation methods rely on segmenting frames individually and using motion information only to segregate moving from static objects and to match segments across frames (Oh et al., 2019; Ventura et al., 2019; Xie et al., 2024). Beyond supervised segmentation, motion information is assumed to be a useful cue for unsupervised learning of segmentation. With few exceptions (Mahendran et al., 2018; Y. Yang et al., 2019), however, object-centric representation learning methods did not use optical flow directly but rather relied on end-to-end reconstruction of videos.

2.3 Connecting human and machine vision

The study of the human visual system and the engineering of machine vision systems have an interconnected history. On the one hand, rebuilding human visual capabilities is a central goal of computer vision. Human vision thus defines many of the problem settings and better understanding the solutions that evolved in biological vision has the potential to inspire computer vision research. On the other hand, the techniques developed in computer vision enable more capable models of human vision that might benefit our understanding of visual perception. The rise of deep neural networks has reinforced interest in both directions and beyond vision, often discussed under the term NeuroAI (Doerig et al., 2023; Zador et al., 2023).

Neuro-inspired machine vision. Research on artificial neural networks has been sparked from insights into the functioning of biological neurons (McCulloch & Pitts, 1943; Rosenblatt, 1958). Likewise, many further milestones during the following decades have been inspired from findings in neuroscience (Fukushima, 1980; Hopfield, 1982). While contemporary research on computer vision is advancing rapidly based on engineering, many aspects in which human vision excels remain unsolved. For example, deep neural networks have been shown to lack robustness to noise (Geirhos et al., 2018) and generalize poorly to novel domains (Zhou et al., 2023). Moreover, while humans are able to continuously learn throughout their lifetime, deep neural networks suffer from *catastrophic forgetting* of previously learned knowledge (French, 1999; McClelland et al., 1995). Finally, the human brain is still orders of magnitude more efficient than state-of-the-art networks (Strubell et al., 2019). Better understanding the solutions implemented in the human visual system might therefore complement engineering-driven computer vision in these respects (Hassabis et al., 2017). In particular, it is commonly assumed that learning a more compositional representation of scenes is necessary to tackle several of these issues (Lake et al., 2017).

DNNs as scientific models. Deep neural networks are able to solve vision tasks on natural scenes while being based on principles inspired by neural networks in the human brain. As such, DNNs have attracted great interest as scientific models of human vision. Deep learning has been particularly successful for building predictive models of neural firing, as it has been shown that DNNs trained on ImageNet outperform previous methods in predicting the firing of neurons in higher visual areas (Cadieu et al., 2014; Yamins & DiCarlo, 2016; Yamins et al., 2014). Moreover, fitting DNNs to neural data directly has enabled better predictive models of neural firing across species (Cadena et al., 2019) that have been shown to captura neural processing in closed loop experiments (Walker et al., 2019).

Beyond applications in neuroscience, DNNs are also tested as perceptual models of human vision (Cichy & Kaiser, 2019; Serre, 2019). Despite achieving remarkable performance for tasks defined on natural scenes, several works have revealed striking differences between deep neural networks such as a bias towards texture (Geirhos et al., 2019), susceptibility to adversarial examples (Szegedy et al., 2014) or lack of generalization to novel viewpoints (Alcorn et al., 2019). Furthermore, direct comparisons between DNNs and human perception are often limited by difficulties due to the experimental paradigms (Funke et al., 2021).

Overall, DNNs are therefore considered as promising models in vision science (Wichmann & Geirhos, 2023), with clear success cases such as neural prediction and task performance that oppose striking differences from human vision. More-

over, prior studies have dominantly focused on core object recognition on images while comparing human and machine perception for other aspects such as motion perception or segmentation has received much less attention. Psychophysics and neuroscience provide a rich history of research on both perceptual phenomena and neural mechanisms related to low-level motion perception while deep learning has been highly successful in enabling high-level perceptual tasks such as segmentation. We see studying the role of motion for structured scene perception therefore as a particularly promising area for connecting human and machine vision.

3 Measuring the Importance of Temporal Features in Video Saliency

This chapter is based on the following publication:

- Matthias Tangemann, Matthias Kümmerer, Thomas S.A. Wallis, and Matthias Bethge. Measuring the Importance of Temporal Features in Video Saliency. ECCV 2020. doi: 10.1007/978-3-030-58604-1_40.

This work was jointly conceptualized by all authors. The selection of datasets and baseline models was collected by M.T. The design of the spatial baseline model was heavily influenced by prior work of M.K., T.W. and M.B. M.B. proposed to curate the meta-benchmark. All experiments and analyses were implemented and executed by M.T., who also wrote the initial draft of the paper. All authors contributed the published version of the paper.

3.1 Motivation

Human vision is an active process. Multiple times per second, we shift our gaze to areas of interest and integrate the visual information into a coherent percept of the surrounding scene. Understanding which factors drive human eye movements is thus essential for a comprehensive understanding of scene perception in humans (Rensink, 2000; Rosenholtz, 1999), and has several applications in computer vision (e.g., Guo and Zhang, 2010).

Models that predict where people look in images have been greatly improved during recent years by adopting deep neural networks (X. Huang et al., 2015; Kruthiventi et al., 2017; Vig et al., 2014). The leading DeepGaze model family (Kümmerer, Theis, & Bethge, 2015; Kümmerer et al., 2017) introduced several new modeling components following a probabilistic approach to gaze prediction. Different from most prior models, DeepGaze predicts a spatial probability distribution and is trained to maximize the likelihood of the observed human gaze. The tendency of humans to look at center of the images is modeled explicitly by fitting a "center bias" distribution and using it as a prior in the model. Combined, these design decision enabled to substantially improve the predictive performance on established benchmarks (Bylinskii et al., 2019; Judd et al., 2012). Furthermore, the probabilistic modeling framework allows to compare model predictions more rigorously by quantifying the model prediction relative to the image-independent center bias in an information theoretic framework. A gold standard model based on inter-observer consistency further provides an estimate of the total information that is predictable. Evaluating models based on information theoretic comparisons

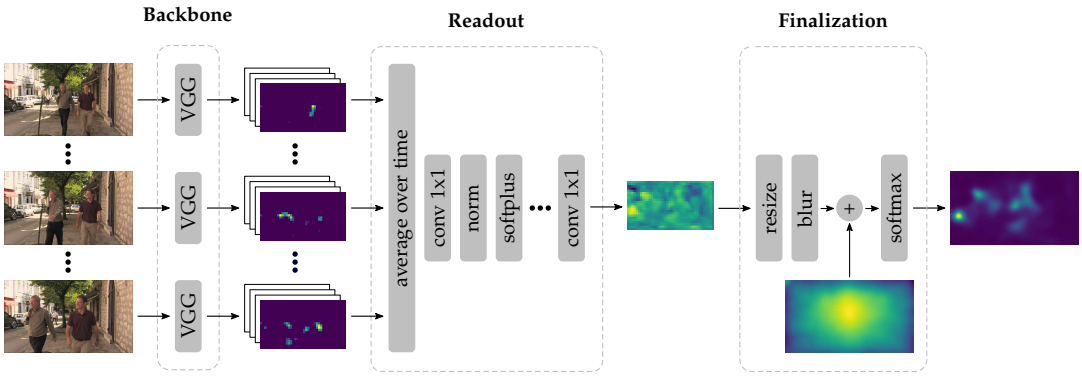

Figure 1: We constructed DeepGazeMR as a baseline model that cannot predict temporal effects by design. DeepGazeMR averages VGG features over time and predicts human gaze using a small readout network and subsequent combination with a learned prior distribution. *Figure used with permission.*

thus allows to quantify model performances more precisely and has been shown to unify previously diverging performance metrics (Kümmerer, Wallis, & Bethge, 2015; Kümmerer et al., 2018).

Several works have extended eye movement prediction to videos. Recently, the LEDOV (L. Jiang et al., 2018; 2019) and DHF1K (W. Wang et al., 2018) datasets have been published which are large enough to support training and benchmarking deep neural networks. Several models have demonstrated the improvements of deep learning over classical methods also for predicting human gaze on video clips (L. Jiang et al., 2018; Lai et al., 2020; Linardos et al., 2019; Min & Corso, 2019; W. Wang et al., 2018). However, these works did not adopt the improved evaluation framework relative to an image-independent baseline and a gold standard model for estimating inter-observer consistency. Thus we do not know how close we are to perfectly predicting human eye movements on videos. Moreover, it has not yet been systematically evaluated to what degree temporal features are necessary to predict human gaze in videos. In our project, we therefore aim to close this gap by a more systematic evaluation of eye movement prediction models on videos.

3.2 Results and synopsis

In this project, we aim to better understand the importance of temporal features to predict where people look when watching videos. To disentangle the influence of spatial and temporal features on human gaze, we constructed a baseline model named *DeepGazeMR* that cannot use temporal information by design (Figure 1). The performance reached by this model therefore is a lower limit on the information

that can be predicted from static appearance alone. We compared the predictions of DeepGazeMR for every frame to a gold standard model as an estimate of inter-subject consistency. Large performance differences between DeepGazeMR and the gold standard reveal situations in which temporal information may by required to predict where people look, which we confirmed by manual case studies.

The overall architecture of DeepGazeMR is based on the state-of-the-art image model DeepGaze II (Kümmerer et al., 2017), but we adapted the architecture of the readout network for our setting. Crucially, features are averaged over time which makes the model invariant to the order of input frames. We trained our model on the LEDOV dataset (L. Jiang et al., 2018), and performed our final analysis on both the LEDOV validation set and the DIEM dataset (Mital et al., 2011). Unfortunately, we could not extend our analysis to the popular DHF1K dataset (W. Wang et al., 2018) due to heavy artifacts in the dataset.

Time-agnostic models improve state-of-the-art for video gaze prediction. We explicitly constructed DeepGazeMR as baseline model that cannot use temporal information. To our surprise, this restricted baseline clearly outperformed previous state-of-the-art models on the LEDOV dataset. The original DeepGaze II model performed slightly better on the DIEM dataset. In both cases, however, all previous video models were outperformed by models that could not use temporal information by design.

Temporal effects are rare in established video datasets. By analyzing the failure cases of our DeepGaze MR baseline, we were able to identify situations in which temporal information is required in order to make an accurate prediction (Figure 2). Our study revealed several clear effects: 1. Interactions between objects shifted gaze to the point of interaction and away from the objects centers. 2. Suddenly appearing elements strongly attract human gaze. 3. Moving object parts attracted more attention. While clear effects exist, these are very rare in the considered datasets. When following the standard practice of averaging model performances over the entire dataset, the ranking is therefore dominated by situations that do not require temporal information. To support the focus on temporal effects, we compiled the situations revealed by our analysis into a novel meta-dataset which we publicly released for the evaluation of future models.

Video models did not learn to predict temporal effects. All prior video models performed poorly when evaluated on our meta-dataset. For the situations considered in our case studies, none of the established video models could predict human gaze better than our spatial baseline. Although being designed and trained on video

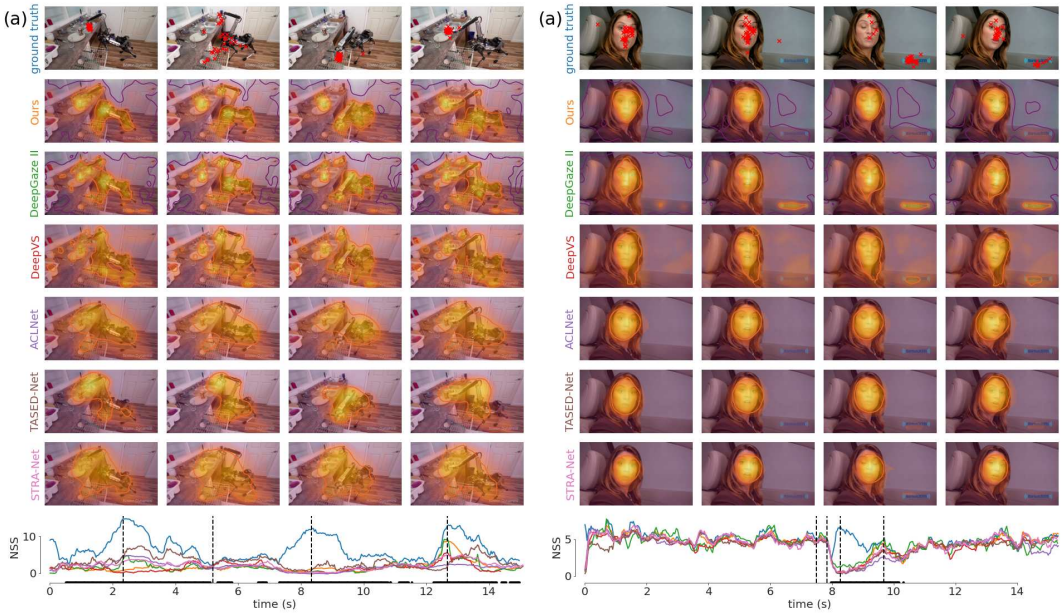

Figure 2: Comparing the gold standard performance to our spatial baseline model allows us to identify situations in which temporal patterns drive human eye movements via case studies. Typical cases include interaction with objects (left) or suddenly appearing objects (right). Prior video saliency models did not improve over our spatial baseline in any of these cases. *Figure used with permission.*

benchmarks, state-of-the-art models thus did not learn to predict the influence of temporal patterns on human gaze.

3.3 Discussion and outlook

Our study revealed the striking failure of video saliency models to capture temporal effects on human gaze. While we found clear cases in which temporal features influence where people look in common datasets, none of the models was able to accurately predict these effects. As such, these models are not yet adequate predictive models of human gaze in dynamic scenes.

We see collecting better training data as the most promising approach for improving video saliency models. All state-of-the art models are based on deep learning and their architectures allow for learning temporal features in principle. Deep neural networks are however known to struggle with learning from imbalanced data (Buda et al., 2018), so that previous models are most likely limited by the scarcity of temporal effects in the training data. While it might be possible to improve the training distribution by non-uniform sampling of the video clips during training,

this will negatively affect the diversity of the training data. Moreover, extending models with explicit mechanisms to capture known temporal effects might improve performance in these cases, but doesn't allow to discover unknown temporal effects in a data driven way. We therefore hypothesize that ultimately it will be necessary collect to training data that contains more temporal effects.

Our study further revealed critical shortcomings of the benchmarks used to evaluate video saliency models. During the last years, deep learning based models have been continuously improving the state-of-the-art performance on established benchmark datasets like LEDOV. Unlike commonly assumed, however, these performance gains did not reflect the better use of temporal information to predict where people look. This kind of benchmarking problem is particularly problematic for deep learning, where benchmarks are used as the main and often only method for model validation due to the inherent difficulty to interpret deep neural networks. Solving the benchmarking problem is however easier to resolve in our case than problems with the training data as it does not require collecting new data. With our meta-benchmark we provide a viable extension to the established benchmarks that explicitly measures the influence of temporal features on human gaze. For any improvements in model mechanisms or training data, it is thus possible to evaluate success in learning temporal effect in comparison to established models (e.g., Kocak et al., 2022).

In summary, our study highlights the importance of data quality and evaluation protocols for the successful application of deep neural networks. Limitations of the available eye movement datasets restrict the scope of current video saliency models for now. Our analysis demonstrates clear temporal effects and suggests clear directions for improvement, so that we see great potential for future deep video saliency models.

4 Unsupervised Object Learning via Common Fate

This chapter is based on the following publication:

- Matthias Tangemann, Steffen Schneider, Julius von Kügelgen, Francesco Locatello, Peter V. Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, Bernhard Schölkopf. Unsupervised Object Learning via Common Fate. CLeaR 2023.

This project was initiated by M.T. and M.B., and led by M.T. The experiments were implemented and executed by M.T. and S.S. in close collaboration: M.T. rendered the Fishbowl dataset and implemented the motion segmentation stage and the architecture of the object and background models. S.S. implemented the scene model. J.v.K. developed the mathematical formalization of the model. The initial draft was written by M.T. and all authors helped to shape the project in internal discussions and contributed to the published version of the paper.

4.1 Motivation

Humans perceive the visual world in a compositional fashion. We perceive scenes as being composed of objects, which can be freely rearranged and recombined to represent infinitely many scenes (e.g., Biederman, 1976; Feldman, 2003; Fodor and Pylyshyn, 1988; Quilty-Dunn et al., 2023). This compositional representation of scenes is commonly believed to be the foundation of humans' capability to quickly adapt to novel situations. In contrast, deep neural networks rival human visual capabilities within the training regime (LeCun et al., 2015) but fail to learn a compositional representation that generalizes systematically (Montero et al., 2021; Schott et al., 2022).

The field of object-centric learning strives to learn a compositional scene representation without human supervision (Goyal & Bengio, 2022; Greff et al., 2020; Lake et al., 2017). Object-centric DNNs are typically based on autoencoders with a latent representation that is factorized into *slots* by design and trained to reconstruct the input scene (Burgess et al., 2019; Locatello et al., 2020; see Yuan et al., 2023 for a review). By adopting a variational autoencoder framework, this approach can be extended to learn a distribution over scenes that allows for the generation of novel scenes (e.g., Engelcke et al., 2021; Eslami et al., 2016). While these methods successfully learn an object-centric representation for simple synthetic scenes composed of colored 3d primitives or digits, they have been shown to fail on more complex scenes which involve textured objects (Karazija et al., 2021; Weis et al., 2021). Moreover, previous approaches typically only learn to represent the visible

object parts which prevents true compositional generalization as objects can only be meaningfully rearranged if they are represented completely.

In this project, we aim to improve object-centric representation learning by taking inspiration from human vision. A range of studies in developmental psychology have allowed to draw conclusions about object learning in infants by using a habituation paradigm to measure the perceptual similarities of scenes at different ages (Kellman & Spelke, 1983; Needham, 1998; Spelke, 1990). These studies indicate that motion and depth cues dominate visual perception in early infancy, while appearance cues are only effective later. It has thus been hypothesized that the early cues like the principle of common fate provide an internal learning signal for learning the appearance of objects. While a line of work exists that extended object-centric models to videos (e.g., Z. He et al., 2019; J. Jiang et al., 2020; Kosiorek et al., 2018), none of these methods attempted to use motion information more explicitly as suggested by developmental psychology. In this project, we aimed to close this gap by implementing an object-centric model that uses the principle of common fate to identify objects and to self-supervise learning an internal object model that allows to complete partial views.

4.2 Results and synopsis

We proposed a multi-stage generative model for compositional scene representation (Figure 3). As the first stage, we used an off-the-shelf motion segmentation model (Keuper et al., 2015) to detect moving regions in the scene. This candidate segmentation is then used as an internal signal to guide the learning of generative models that represent backgrounds and objects, respectively. For both the object and background models, we adopt a variational autoencoder framework but limit the reconstruction loss to the regions that are predicted relevant by the motion segmentation. Finally, the object and background models are combined into a generative, compositional scene model that learns a distribution over object counts and arrangements.

To evaluate our model, we created a synthetic dataset positioned between the simplistic datasets used by previous works and natural scenes. Additionally to the input videos and ground truth segmentation, we rendered the unoccluded objects for all evaluation videos in order to test the model's capability for amodal completion.

Overall, we found that our infant vision inspired approach to object learning improved over previous object-centric models in several ways:

The object model learns to represent complete objects. During training, the motion segmentation stage only yields partial mask for the occluded objects. Nevertheless, our object model successfully learned to reconstruct complete objects

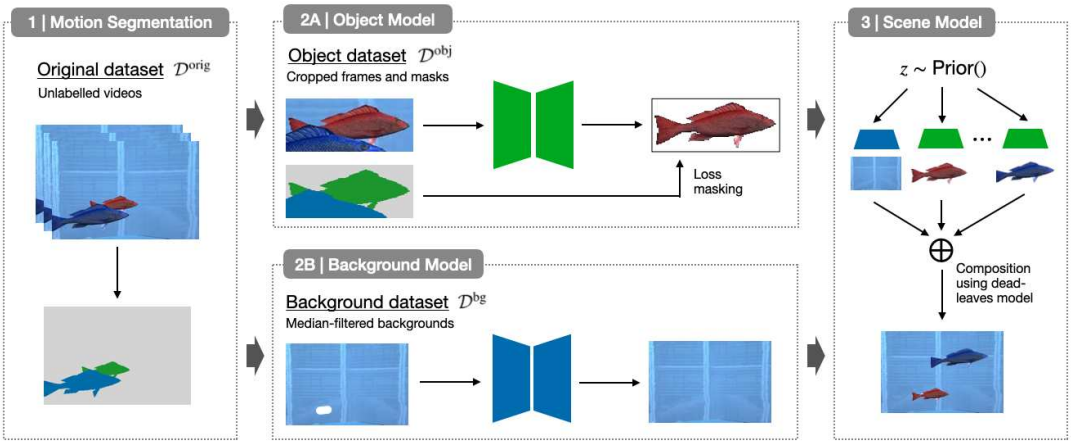

Figure 3: We follow a multi-stage modeling approach inspired by human vision: Moving region are detected in the first stage and used to guide the learning of internal background and object models. The scene model combines the background and object models and allows for the controlled sampling of novel scenes.

across various occlusion levels. This result indicates that learning complete objects is the simplest internal model that explains the various partially occluded object views during training, without requiring explicit supervision on the hidden parts. Furthermore, we found that the ability of the model to complete objects critically depends on the quality of the motion segmentation: Simulating a perfect motion segmentation using ground truth data revealed a substantial gap to our model, especially for highly occluded objects.

Our model scales to more realistic inputs. A model comparison confirmed previous results which showed that existing object-centric models struggled to learn objects for more realistic, textured scenes. In particular, we observed some models to learn a viable scene reconstruction without a correct segmentation, and failures to generate novel scenes. Our approach outperformed previous approaches both in terms of compositional representation and visual quality.

Our model allows for controlled scene generation beyond the training distribution. The scene model in our approach uses the object and background models as basic building blocks, and only captures high-level scene parameters like the distribution of object counts and locations. This compositional sampling process allows for controlled interventions, such as moving or exchanging individual objects in a scene (Figure 4). Moreover, it is straightforward to meaningfully generalize beyond the

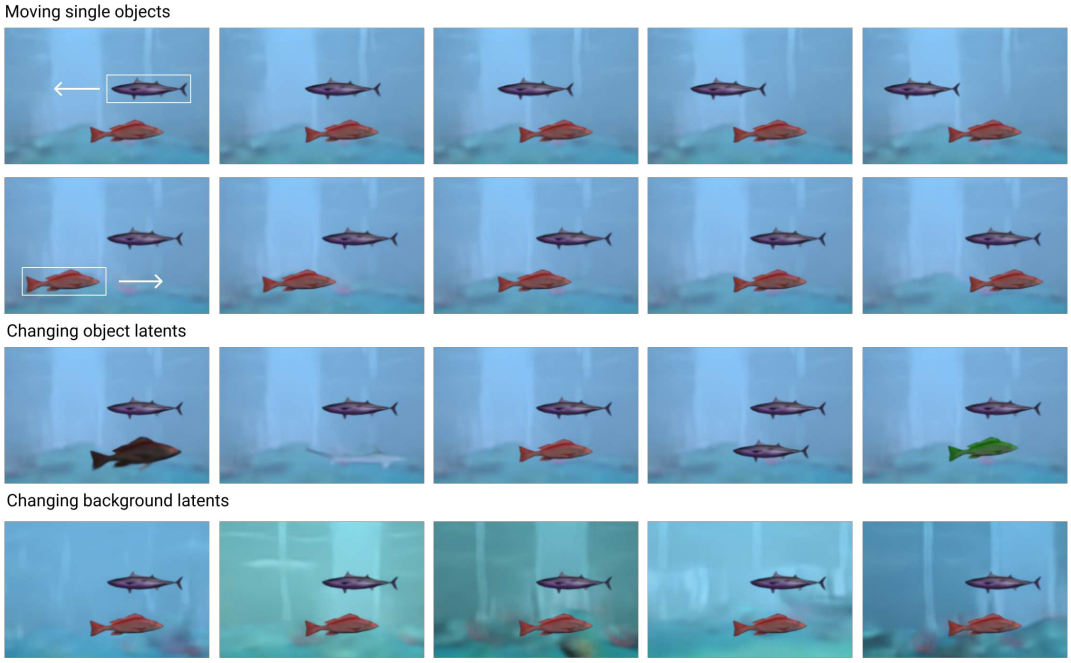

Figure 4: Our models represents scenes in a compositional way, and thus allows to intervene on individual objects.

input data and create, for example, scenes with more objects as have been observed during training.

Our model can quickly adapt to novel scene statistics. As the scene model only captures high-level parameters, it can be quickly adapted to novel scene statistics in a few-shot fashion. For example, given a scene in which objects only appear in certain locations but never in others, the model can be fitted to this new scene statics using only a single example video.

4.3 Discussion and outlook

In this project, we have shown that insights from developmental psychology can be successfully transferred to a computer vision setting. Using common fate as an internal learning signal allowed us to scale object-centric representation learning to more realistic, textured inputs and to represent scenes in a more physically plausible way. Following our work, several other groups have further improved the quality and segmentation performance of motion-based object centric models (Bao et al., 2022; 2023; Chen et al., 2022; Choudhury et al., 2022; Karazija et al., 2022).

While the use of motion information is a promising path towards scalable object-centric learning, alternative approaches have enabled improvements more quickly during the years following our initial publication. Large-scale self-supervised models like DINO (Caron et al., 2021; Oquab et al., 2024) have been demonstrated to learn a visual representation that successfully transfers to many downstream tasks. While not build as compositional models, applying standard object-centric learning or clustering methods to their internal representation has resulted in scalable object discovery (Seitzer et al., 2023; X. Wang et al., 2024; X. Wang et al., 2023; Zadaianchuk et al., 2023). In the same way, non-compositional image generation models like Stable Diffusion (Rombach et al., 2022) have been shown to learn a good basis for object discovery as well (Couairon et al., 2024). Moreover, the Segment Anything Model SAM has been a breakthrough in image segmentation which has been enabled by supervised learning on carefully engineered training data at unprecedented scale (Kirillov et al., 2023). The success of SAM challenges the necessity of human-like, unsupervised segmentation learning for practical applications.

Beyond these advancements in computer vision, progress in AI has been mainly driven by large language models (Brown et al., 2020; Touvron et al., 2023) and extensions of these models to support vision (Liu et al., 2023; OpenAI et al., 2024). While evidence from cognitive science shows that a compositional representation precedes the emergence of language (Fedorenko et al., 2024), the available mass of text on the internet represents a unique shortcut for engineering. After all, language is an inherently compositional system that reflects human perception of the world, so that training on large amounts of text is expected to foster learning a compositional representation.

While classical object-centric learning might therefore seem less relevant now, many of the underlying questions are not yet solved even by the most sophisticated models. Several studies have pointed out striking failures of large multimodal models to capture the fundamental spatial and physical regularities that govern our visual environment (Bonnen et al., 2024; Ramakrishnan et al., 2024). Moreover, object-level control is still an active area of research in state-of-the-art generative models (Wan et al., 2024). Machine vision thus does not yet represent scenes as a coherent spatial arrangement of three-dimensional objects and their physical relations—which infants learn at very young age. Our model's capability for amodal completation by the representation of complete objects is a step in this direction at a small scale. Insights from developmental psychology and careful comparison to human perception have the potential to further guide the improvement of spatial reasoning in large scale models.

5 Object segmentation from common fate: Motion energy processing enables human-like zero-shot generalization to random dot stimuli

This chapter is based on the following publication:

- Matthias Tangemann, Matthias Kümmerer, Matthias Bethge. Object segmentation from *common fate*: Motion energy processing enables human-like zero-shot generalization to random dot stimuli. NeurIPS 2024.

This work was initiated and led by M.T. All authors supported the design of this study. The model was designed by M.T., who also implemented and executed all experiments. The paper was written by M.T., with valuable contributions from M.B.

5.1 Motivation

The principle of common fate, which states that elements that move together belong together, is an essential cue utilized by the human visual system for perceptual organization (Wertheimer, 1912). In the previous project, we took inspiration from developmental psychology (Spelke, 1990) to develop a model that learns a compositional scene representation based on motion information. A critical factor for learning a good scene representation was the quality of the motion segmentation, which we assumed as given in the previous project by using an off-the-shelf approach from computer vision. While this assumption is plausible from the perspective of developmental psychology (Arterberry & Kellman, 2016), we don't have a good understanding yet of the mechanisms used by the human brain to segment moving objects in real world scenes (Nishida et al., 2018). In this project, we therefore aimed to find a good model of moving object perception in humans and tested a wide range of computer vision models for alignment with human perception.

A crucial property of the principle of common fate in human vision is the ability to detect moving objects irrespective of their visual appearance. This independence from appearance is essential when using common fate as a learning signal for object appearance in infancy, but also supports robust visual perception in cases where appearance information is unreliable during adulthood. A striking example are camouflaged animals, which are hard to detect based on their visual appearance but easily spotted by humans when moving. In controlled vision science experiments, many previous works have used random dot stimuli in order to study motion processing in isolation (e.g., Johansson, 1973, Newsome and Pare, 1988). Based on a natural video, it is possible to construct corresponding random dot stimuli

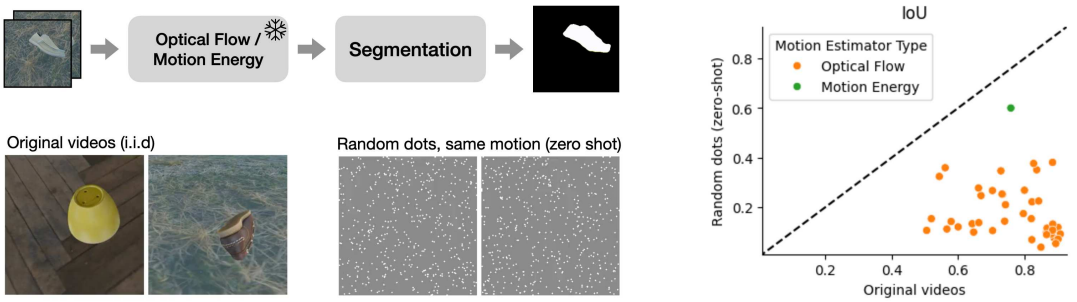

Figure 5: We compared optical flow methods and a neuroscience inspired motion energy model for the downstream task of appearance free motion segmentation. For each motion estimation model, we trained a segmentation network on vidoes of moving objects. We tested all models on i.i.d. evaluation videos and corresponding random dot videos which preserve the motion but erase the appearance information of the original videos. The neuroscience inspired motion energy model substantially outperforms all optical flow models in terms of generalization to random dots.

that preserve the motion information of the original video but remove all appearance information. Nevertheless, humans have been shown to reliably recognize moving objects using motion information alone without prior exposure to random dot patterns (Robert et al., 2023). In this project, we therefore evaluated which computational models are aligned with human visual perception and support the segmentation of random dot patterns.

5.2 Results and synopsis

Motion segmentation is typically approached as a two stage process. In the first stage, a dense motion field is estimated for the input video which is then used as input for the segmentation model. Using two-frame optical flow estimators for the first stage, this approach has enabled impressive performance gains on standard computer vision benchmarks (Tokmakov et al., 2019; Xie et al., 2022). For generalization to random dot stimuli, the motion estimation stage should be ideally invariant to changes in texture. We therefore focused on testing a large range of optical flow models, while keeping the segmentation stage of the model fixed. Additionally to standard optical flow models, we tested a neuroscience inspired motion energy model (Simoncelli & Heeger, 1998). While originally developed to explain the firing rates of neurons involved in motion perception in visual area MT, this model can be applied to videos in a dense fashion and used as the first stage of a motion segmentation approach.

We created a benchmark dataset for testing generalization to random dot stimuli (Figure 5). Using the Kubric simulator (Greff et al., 2022), we created videos of scanned objects that are moving relative to the background. For each motion estimation model, we used a subset of the videos to train the segmentation stage for figure-ground segmentation. Critically, the training set only contained the unmodified, textured videos. For evaluation, we applied the models to both the original videos and the corresponding random dot videos from a separate validation subset. The motion in both conditions is the same by construction, such that appearance-agnostic models should reach the same performance for both conditions. However, our analysis revealed a large difference between the computer vision and neuroscience inspired models:

State-of-the-art motion segmentation does not generalize to random dots. Overall, the computer vision approaches excel on the i.i.d. test set and allow for an almost perfect segmentation of the original videos. However, all models performed substantially worse for the random dot stimuli. This performance gap was particularly the large for the most recent optical flow models that perform best in standard benchmarks (e.g. Shi et al., 2023), with GMFlow (Xu et al., 2022) being a notable exception.

A classic motion energy model enables generalization to random dots. Despite predating deep learning and not being designed for dense motion estimation in a computer vision setting, the motion energy model of Simoncelli and Heeger (1998) can be successfully used in a modern motion segmentation pipeline. The performance on the original videos does not match the performance enabled by state-of-the-art optical flow models, but outperforms several of the early deep learning based methods. Most importantly, however, the motion energy model substantially outperforms all optical flow models in terms of generalization to random dots by a huge margin.

While it has been shown that humans can classify objects in moving random dot patterns (Robert et al., 2023), it has not been systematically tested before how well humans can segment objects. We therefore performed a psychophysical shape identification task to directly compare humans and machines (Figure 6). In each trial, a random dot video was shown to the participants. After the video stopped playing, the participants had to select one of two alternative shapes which they observed in the video. We applied the segmentation algorithms to the same task, by using them to segment the random dot video and predict the shape alternative with the larger overlap as measured by IoU.

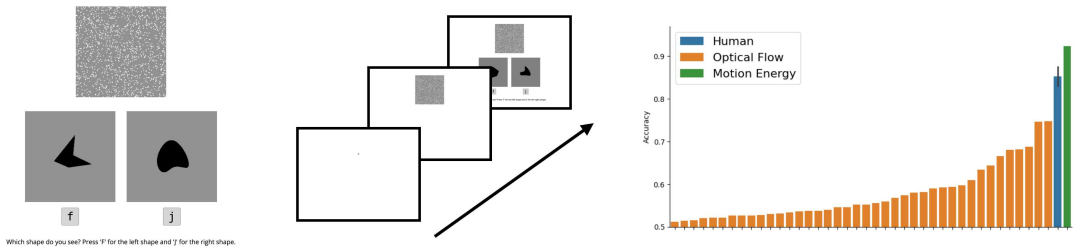

Figure 6: We directly compared human and machines for random dot motion segmentation in a psychophysical experiment. Participants had to identify a shape in a random dot stimulus given two options that where only shown after the video stopped playing. Only the motion energy model, but not state-of-the-art optical flow models, enabled reaching human performance.

Only the motion energy model matches human perception in a direct comparison. The human participants performed well in the task, reaching an accuracy of 80% correct on average. While some of the optical flow based motion segmentation models reached a non-trivial performance on this task, they fell behind human capabilities by a substantial margin. The motion energy based motion segmentation model was the only model to match human performance and performed similarly as the best participants in our study.

5.3 Discussion and outlook

Deep neural networks have been described as "promising—but not yet adequate" models of human vision (Wichmann & Geirhos, 2023). Much research has been comparing human perception and computer vision for the task of core object recognition. Successes such as high task performance on natural scenes and prediction of neural firing rates (Yamins et al., 2014) are accompanied by striking differences to human vision (Geirhos et al., 2020). Motion perception is an essential aspect of human vision where computer vision reached high task performance, but only few studies compared humans and machines in this setting (Y.-H. Yang et al., 2023) and our work is the first to compare humans and machines for the task of segmenting moving objects. While humans are able to apply the Gestalt principle of common fate independently of appearance, our analysis revealed a striking failure of state-of-the-art optical flow models in this case. Our study therefore makes an important contribution towards a comprehensive evaluation of DNNs as models of human vision and complement other works on core object recognition that point out differences between human and machine perception despite high task performance.

We see motion perception is particularly promising for closing the gap between human perception and computer vision since the low-level mechanisms for mo-

tion detection have been extensively studied in neuroscience. Starting with the model proposed by Hassenstein and Reichardt (1956), numerous models of motion perception have been proposed and extensively validated using psychophysical experiments and neural recordings (Adelson & Movshon, 1982; Simoncelli & Heeger, 1998). Our study, as well as a parallel work by another group (Z. Sun et al., 2023), demonstrate the possibility to transfer these approaches to modern computer vision. Despite receiving much less attention than mainstream computer vision approaches, the gap in task performance is surprisingly small. Therefore we hypothesize that models which unify task performance with similarity to human perception are in reach. Most importantly, these models have the potential to realize the promise of deep neural networks as scientific models to bridge neural mechanisms and human behavior (Doerig et al., 2023). Our work is an important step in this direction, as it revealed a compelling link between the mechanisms of cortical motion processing and the Gestalt principle of common fate.

Finally, our work might offer some inspiration to build better computer vision systems for motion perception. The high performance of contemporary optical flow networks comes at the prize of large models with high computational cost while the motion energy energy model has several orders of magnitude fewer parameters but offers promising performance. Inspiration from neuroscience might therefore support applications where efficiency is critical, such as fast-moving mobile robots where standard computer vision methods are not adequate (Falanga et al., 2020). Moreover, established optical computer vision methods have been shown to be highly susceptible to noise (Geirhos et al., 2018), while the motion energy model was highly robust to appearance changes in our setting. Studies like ours, that are positioned at the intersection of neuroscience and machine learning, might therefore directly contribute the advancement of computer vision applications in these cases.

6 Discussion

Humans excel at interpreting complex visual scenes by structuring raw sensory input into high-level, object-centric representations. Motion is considered a key cue in this process, influencing the development of visual perception, attention and the grouping of scene elements. While extensive research has provided insights into these aspects using highly controlled, artificial stimuli, recent advances in deep learning have enabled models that predict human behavior and neural responses in more naturalistic settings. However, prior studies have primarily focused on modeling object recognition in static images. In this thesis, we extend this approach to examine the role of motion in different aspects of scene perception.

- Chapter 3 investigates *motion as an attention cue*. Using eye-tracking data from video datasets, we identified clear effects of motion on gaze behavior in natural scenes. However, we also found that such effects are rare in existing benchmarks. To address this, we introduced a meta-benchmark that consolidates relevant cases across datasets.

- Chapter 4 explores *motion as a learning cue*. We developed a model that learns to disentangle objects and background representations based on motion information. This approach captures key aspects of human scene perception, including the ability to infer occluded object parts and generate novel scenes beyond the training distribution.

- Chapter 5 examines *motion as a segmentation cue*. Humans can segment objects from the background in random-dot motion stimuli without prior experience. We showed that a motion energy approach, unlike traditional optical flow algorithms, enables this appearance-free motion segmentation and thus provide a link between cortical mechanisms of motion processing and the Gestalt principle of common fate.

This work bridges neuroscience, cognitive science, and artificial intelligence. A central aspect of this emerging field of NeuroAI is the two way interaction between those fields: Deep neural networks are used as scientific models with the aim to improve our understanding of human vision, and human vision serves as inspiration to build more capable artificial vision systems (Momennejad, 2022). In this spirit, the following discussion first focuses on our findings regarding how the human visual system utilizes motion. We then consider implications for artificial vision systems before concluding with an outlook on research at the intersection of human and machine vision.

6.1 Human vision

Motion has long been recognized as a central cue in human vision, playing a crucial role in directing attention (Rosenholtz, 1999), learning object representations (Spelke, 1990), and structuring visual scenes (Wertheimer, 1912). The studies presented in this thesis reinforce the significance of motion in all these contexts. By leveraging deep learning, we extend this line of research in several important ways, providing new insights into how motion influences human vision in naturalistic settings and by advancing the available tools to study these processes.

Classical studies on human vision have often relied on simple and controlled stimuli, such as moving dots and gratings (Simoncelli & Heeger, 1998). While these studies have been invaluable in identifying fundamental mechanisms, real-world vision involves vastly more complex inputs. Deep neural networks have demonstrated great task performance on natural inputs, and as such are considered a promising method for scaling models of human vision (Wichmann & Geirhos, 2023). Our work contributes to realizing this potential and helps bridging the gap between controlled experiments and real-world perception. In Chapter 3, we showed that deep learning can be used to disentangle the influence of static and temporal features in video saliency. By applying this method to an established eye-movement dataset, we systematically identified cases where motion influences human attention in a data-driven way. In Chapter 5, we combined a classical neuroscience-inspired model of motion perception (Simoncelli & Heeger, 1998) with deep neural networks into a figure-ground segmentation model for naturalistic videos. The interpretation of motion patterns in terms of high-level scene structure has been recognized as an important open question in motion perception (Nishida et al., 2018). By connecting low-level mechanisms of cortical motion processing to the Gestalt principle of common fate in naturalistic scenes, we made an important contribution to closing this gap. Across projects, our work therefore advances a more holistic understanding of human vision under more natural conditions.

While deep neural networks have generally been praised for their high task performance, their usefulness as models for vision science has been questioned (Bowers et al., 2023). In particular, critics argue that DNNs do not contribute to our understanding of human vision because they are inherently difficult to interpret. Our studies make relevant contributions to this debate. On the one hand, our work on benchmarking video saliency models in Chapter 3 highlights the risks of the benchmark-driven approach in contemporary machine learning. Discrepancies between the intention of a benchmark and factors influencing good performance may lead to false conclusions that are difficult to detect. On the other hand, our work shows that these problems can be addressed through careful

analysis. By using a restricted baseline model, we were able to reveal benchmark shortcomings and provide a subset that allows for more rigorous testing of video saliency models. Vision science has a long tradition of designing experiments and controls to support reliable conclusions about human vision. When approaching the design and evaluation of benchmarks with the same rigor, we see deep learning as valuable addition to more established research methods. Furthermore, our work on motion energy segmentation in Chapter 5 demonstrates that it is possible to combine classical, interpretable models from computational neuroscience with deep learning to unify interpretability and task performance. We see great potential in such an "interpretability-by-design" approach for future research. In summary, our work demonstrates that deep learning can be a useful tool to advance our understanding of human vision, but, like any research method, it requires careful experimental design and an awareness of its limitations.

Beyond scaling towards natural scenes, deep learning enables more specific formulation of hypotheses. Developing an image-computable model requires specifying a connection between low-level mechanisms and observed behavior—a link still missing for many aspects of human vision (Doerig et al., 2023). Motion has long been hypothesized as a key cue that infants use to decompose scenes into objects (Spelke, 1990). Similarly, since the early works on Gestalt psychology, common fate has been described as a fundamental principle for perceptual organization in the adult visual system (Wertheimer, 1912). For both cases, we contributed to a more precise understanding of the underlying mechanisms by providing image-computable models that are able to make predictions for arbitrary videos (Chapters 4 and 5). The predictions of our models align with human perception in several ways, such as amodal completion and the ability to segment random dot stimuli. Future experiments may reveal differences between our models and human perception, leading to refinements and improvements. In the spirit of the neuroconnectionist research program (Doerig et al., 2023), we see the models proposed in this work as executable hypotheses that advance our understanding of the connection between mechanisms and behavior. In this way, our work represents a significant advancement in the study of motion perception in humans.

Deep learning has substantially improved models of visual perception during the last years. However, several challenges remain in building a complete computational model of scene perception in human vision (Serre, 2019; Wichmann & Geirhos, 2023). First, we don't have a complete understanding of the similarities and differences between human and machine vision yet. While many studies evaluated deep neural networks as a model for core object recognition, other aspects like motion or depth perception have just begun to be explored. Second, despite human-like task performance offered by DNNs, more detailed analyses reveal differences between

human perception and computer vision in many cases, for example in terms of error patterns or robustness (Geirhos et al., 2020). Third, we are lacking methods to translate the predictive performance of DNNs into tangible explanations of human perception (e.g. Saxe et al., 2021). This includes both tools for interpreting the learning outcome of deep neural networks as well as tools for integrating classical, interpretable methods with the power of deep learning.

In summary, this thesis extends research on motion perception in human vision by leveraging deep learning to analyze motion's role in gaze behavior, object learning, and segmentation. By pushing towards naturalistic stimuli, providing image-computable models as executable hypotheses, and linking mechanisms with behavior, we contribute to a more comprehensive understanding of motion in human vision.

6.2 Machine vision

Throughout the history of artificial intelligence, inspiration from natural intelligence has played a crucial role in advancing computational models (e.g., Fukushima, 1980; Rosenblatt, 1958). Whether this approach remains valuable is an open debate (Hassabis et al., 2017; Zador et al., 2023). On the one hand, the human brain provides a proven implementation for general intelligence and understanding the underlying mechanisms would certainly be valuable for building an artificial models. However, the mechanisms of the brain are most likely not the only possible implementation. Recent developments, particularly in large language models, suggest a growing divergence between the solutions driving human intelligence and artificial intelligence. Our work contributes to this discussion by offering concrete examples of both potential successes and limitations of biologically inspired AI.

One of the most direct instances of inspiration from human vision in our work is found in Chapter 4, where insights from from infant learning helped to improve object-centric learning. While this approach was beneficial within the specific context of object-centric representations, it remains an open question whether object-centric learning itself is the right approach for artificial vision. Models such as the Segment Anything Model SAM (Kirillov et al., 2023) demonstrate the strengths of supervised learning, while representation learning approaches such as DINO (Caron et al., 2021; Oquab et al., 2024) and CLIP (Radford et al., 2021) show that non-compositional methods can be highly effective. These models, however, diverge significantly from the way humans learn, reinforcing the question of whether biologically inspired approaches are necessary for achieving artificial vision.

Despite these differences, there are reasons to seek inspiration from human vision, particularly in designing interfaces between AI and humans. AI systems will

increasingly interact with people, and their usability could benefit from making errors that resemble human mistakes. This could also have implications for AI safety, as systems that fail in ways understandable to humans may be easier to predict and control (Mineault et al., 2024). Moreover, studies on human vision can guide AI research by highlighting critical areas where artificial models diverge from human perception. In Chapter 5, we have demonstrated striking differences between human and machine motion perception in terms of robustness to appearance changes. Other authors have shown, for example, substantial differences in humans and machines regarding aspects of spatial vision (Bonnen et al., 2024; Ramakrishnan et al., 2024). By identifying such discrepancies, research on human vision can provide insights into improving artificial vision systems—regardless of whether the solutions come from biological inspiration or engineering.

We see improving the efficiency of dynamic scene perception as a particularly promising area where research on human vision might inspire artificial vision systems. Typical state-of-the-art models treat motion as an extension of image-based approaches rather than exploiting unique temporal properties of dynamic scenes. For instance, optical flow networks often rely on only two frames, treating motion as an image-matching problem rather than a true temporal inference task (Xu et al., 2022). Similarly, methods for object tracking mostly follow a *tracking-by-detection* approach and focus most of their computation on single-frame analysis, with only minimal processing devoted to frame-to-frame associations (e.g., Lv et al., 2024). Likewise, leading motion segmentation and video object segmentation models prioritize segmenting individual frames before integrating them over time (Ravi et al., 2024; Xie et al., 2024). This results in computational inefficiencies that make deep learning models difficult to deploy in real-time applications that require high frame rates. In contrast, biological vision systems, from insects to primates, exhibit highly efficient motion processing mechanisms. Using a neuro-inspired motion energy model in Chapter 5 has shown promising performance in direct comparison with recent optical flow models, despite having several orders of magnitude fewer parameters. We therefore see significant potential for further improving computer vision systems by drawing from these biological strategies.

6.3 Outlook

The fields of neuroscience and artificial intelligence have a common history, and the emerging field of NeuroAI renews the interest in bidirectional inspiration between those fields (Doerig et al., 2023; Momennejad, 2022; Zador et al., 2023). The projects in this thesis have contributed in both directions by using deep neural networks as models of motion perception in humans and using insights from neuroscience

and developmental psychology to build more capable machine learning models. In the future, we see great potential for NeuroAI to advance our understanding of intelligence. Benchmarks inspired from human capabilities may provide guidance towards more capable and more reliable artificial vision systems, and mechanisms evolved in biological brains may provide directions for increasing efficiency. In particular, we expect substantial progress in understanding the human visual system by using deep neural networks as scientific models. Deep learning is highly successful in solving vision tasks for natural stimuli. While studies have shown striking difference between contemporary DNNs and human perception, many works, including our own, have also shown paths towards narrowing this gap in the future (Geirhos et al., 2021; Jaini et al., 2024). Building a complete model of human visual perception therefore seems possible, and we expect that working towards this goal will significantly advance vision science.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299. https://doi.org/10.1364/JOSAA.2.000284

Adelson, E. H., & Movshon, J. A. (1982). Phenomenal coherence of moving visual patterns. *Nature*, *300*(5892), 523–525. https://doi.org/10.1038/300523a0

Albright, T. D. (1984). Direction and orientation selectivity of neurons in visual area MT of the macaque. *Journal of Neurophysiology*, *52*(6), 1106–1130. https://doi.org/10.1152/jn.1984.52.6.1106

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & Nguyen, A. (2019). Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4845–4854.

Arterberry, M. E., & Kellman, P. J. (2016, June 9). *Development of Perception in Infancy: The Cradle of Knowledge Revisited* (2nd edition). Oxford University Press.

Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.-X., Gaidon, A., & Hebert, M. (2022). Discovering Objects That Can Move. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11789–11798.

Bao, Z., Tokmakov, P., Wang, Y.-X., Gaidon, A., & Hebert, M. (2023). Object Discovery From Motion-Guided Tokens. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22972–22981.

Barlow, H. B., & Levick, W. R. (1965). The mechanism of directionally selective units in rabbit's retina. *The Journal of Physiology*, *178*(3), 477–504. https://doi.org/10.1113/jphysiol.1965.sp007638

Barlow, H. B., & Hill, R. M. (1963). Selective Sensitivity to Direction of Movement in Ganglion Cells of the Rabbit Retina. *Science*, *139*(3553), 412–414. https://doi.org/10.1126/science.139.3553.412

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018, October 17). *Relational inductive biases, deep learning, and graph networks*. arXiv: 1806.01261. https://doi.org/10.48550/arXiv.1806.01261

Bideau, P., RoyChowdhury, A., Menon, R. R., & Learned-Miller, E. (2018). The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 508–517.

Biederman, I. (1976). On processing information from a glance at a scene: Some implications for a syntax and semantics of visual processing. *Proceedings of the ACM/SIGGRAPH Workshop on User-oriented Design of Interactive Graphics Systems*, 75–88. https://doi.org/10.1145/1024273.1024283

Black, M., & Anandan, P. (1993). A framework for the robust estimation of optical flow. *1993 (4th) International Conference on Computer Vision*, 231–236. https://doi.org/10.1109/ICCV.1993.378214

Bonnen, T., Fu, S., Bai, Y., O'Connell, T., Friedman, Y., Kanwisher, N., Tenenbaum, J. B., & Efros, A. A. (2024, September 10). *Evaluating Multiview Object Consistency in Humans and Image Models*. arXiv: 2409.05862. https://doi.org/10.48550/arXiv.2409.05862

Borst, A. (2000). Models of motion detection. *Nature Neuroscience*, *3*(11), 1168–1168. https://doi.org/10.1038/81435

Borst, A., & Egelhaaf, M. (1989). Principles of visual motion detection. *Trends in Neurosciences*, *12*(8), 297–306. https://doi.org/10.1016/0166-2236(89)90010-6

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385. https://doi.org/10.1017/S0140525X22002813

Braddick, O. (1974). A short-range process in apparent motion. *Vision Research*, *14*(7), 519–527. https://doi.org/10.1016/0042-6989(74)90041-8

Braddick, O. (1993). Segmentation versus integration in visual motion processing. *Trends in Neurosciences*, *16*(7), 263–268. https://doi.org/10.1016/0166-2236(93)90179-P

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020, July 22). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs]. https://doi.org/10.48550/arXiv.2005.14165

Brox, T., & Malik, J. (2010, September). Object Segmentation by Long Term Analysis of Point Trajectories. In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), *Computer Vision – ECCV 2010* (pp. 282–295, Vol. 6315). Springer. https://doi.org/10.1007/978-3-642-15555-0_21

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., & Lerchner, A. (2019, January 22). *MONet: Unsupervised Scene Decomposition and Representation*. arXiv: 1901.11390. https://doi.org/10.48550/arXiv.1901.11390

Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012, October). A Naturalistic Open Source Movie for Optical Flow Evaluation. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012* (pp. 611–625, Vol. 7577). Springer. https://doi.org/10.1007/978-3-642-33783-3_44

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2019). What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLOS Computational Biology*, *15*(4), e1006897. https://doi.org/10.1371/journal.pcbi.1006897

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-End Object Detection with Transformers. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 213–229). Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_13

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

Cavanagh, P. (1991). Short-range vs Long-range motion: Not a valid distinction. *Spatial Vision*, *5*(4), 303–309. https://doi.org/10.1163/156856891X00065

Chen, H., Venkatesh, R., Friedman, Y., Wu, J., Tenenbaum, J. B., Yamins, D. L. K., & Bear, D. M. (2022, October). Unsupervised Segmentation in Real-World Images via Spelke Object Inference. In S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, & T. Hassner (Eds.), *Computer Vision – ECCV 2022* (pp. 719–735). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19818-2_41

Choudhury, S., Karazija, L., Laina, I., Vedaldi, A., & Rupprecht, C. (2022). Guess What Moves: Unsupervised Video and Image Segmentation by Anticipating

Motion. *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*.

Chubb, C., & Sperling, G. (1988). Drift-balanced random stimuli: A general basis for studying non-Fourier motion perception. *Journal of the Optical Society of America A*, *5*(11), 1986–2007. https://doi.org/10.1364/JOSAA.5.001986

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Couairon, P., Shukor, M., Haugeard, J.-E., Cord, M., & Thome, N. (2024, October 5). *DiffCut: Catalyzing Zero-Shot Semantic Segmentation with Diffusion Features and Recursive Normalized Cut*. arXiv: 2406.02842. https://doi.org/10.48550/arXiv.2406.02842

Curcio, C. A., Sloan, K. R., Kalina, R. E., & Hendrickson, A. E. (1990). Human photoreceptor topography. *Journal of Comparative Neurology*, *292*(4), 497–523. https://doi.org/10.1002/cne.902920402

Dave, A., Tokmakov, P., & Ramanan, D. (2019). Towards Segmenting Anything That Moves. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Dayan, P., & Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Doerschner, K., Fleming, R. W., Yilmaz, O., Schrater, P. R., Hartung, B., & Kersten, D. (2011). Visual Motion and the Perception of Surface Material. *Current Biology*, *21*(23), 2010–2016. https://doi.org/10.1016/j.cub.2011.10.036

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., & Brox, T. (2015). FlowNet: Learning Optical Flow With Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2758–2766.

Duffy, C. J., & Wurtz, R. H. (1991). Sensitivity of MST neurons to optic flow stimuli. I. A continuum of response selectivity to large-field stimuli. *Journal of Neurophysiology*, *65*(6), 1329–1345. https://doi.org/10.1152/jn.1991.65.6.1329

Engelcke, M., Parker Jones, O., & Posner, I. (2021, December). GENESIS-V2: Inferring Unordered Object Representations without Iterative Refinement. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. Wortman Vaughan (Eds.),

Advances in Neural Information Processing Systems (pp. 8085–8094, Vol. 34). Curran Associates, Inc.

Eslami, S. M. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., kavukcuoglu, k., & Hinton, G. E. (2016, December). Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Exner, S. (1875). Experimentelle Untersuchung der einfachsten psychischen Processe. *Archiv für die gesamte Physiologie des Menschen und der Tiere*, *11*(1), 403–432. https://doi.org/10.1007/BF01659311

Falanga, D., Kleber, K., & Scaramuzza, D. (2020). Dynamic obstacle avoidance for quadrotors with event cameras. *Science Robotics*, *5*(40), eaaz9712. https://doi.org/10.1126/scirobotics.aaz9712

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. F. (2024). Language is primarily a tool for communication rather than thought. *Nature*, *630*(8017), 575–586. https://doi.org/10.1038/s41586-024-07522-w

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252–256. https://doi.org/10.1016/S1364-6613(03)00111-6

Fennema, C. L., & Thompson, W. B. (1979). Velocity determination in scenes containing several moving objects. *Computer Graphics and Image Processing*, *9*(4), 301–315. https://doi.org/10.1016/0146-664X(79)90097-2

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, *3*(4), 128–135. https://doi.org/10.1016/S1364-6613(99)01294-2

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S. A., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, *21*(3), 16. https://doi.org/10.1167/jov.21.3.16

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature*

Machine Intelligence, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021, December). Partial success in closing the gap between human and machine vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (pp. 23885–23899, Vol. 34). Curran Associates, Inc.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th International Conference on Learning Representations (ICLR) 2019*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018, December). Generalisation in humans and deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc.

Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 580–587.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068. https://doi.org/10.1098/rspa.2021.0068

Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D. J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.-T. (, Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., … Tagliasacchi, A. (2022). Kubric: A Scalable Dataset Generator. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3749–3761.

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020, December 9). *On the Binding Problem in Artificial Neural Networks*. arXiv: 2012.05208 [cs]. https://doi.org/10.48550/arXiv.2012.05208

Guo, C., & Zhang, L. (2010). A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. *IEEE Transactions on Image Processing*, *19*(1), 185–198. https://doi.org/10.1109/TIP.2009.2030969

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hassenstein, B., & Reichardt, W. (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. *Zeitschrift für Naturforschung B*, *11*(9-10), 513–524. https://doi.org/10.1515/znb-1956-9-1004

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

He, Z., Li, J., Liu, D., He, H., & Barber, D. (2019). Tracking by Animation: Unsupervised Learning of Multi-Object Attentive Trackers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1318–1327.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Horn, B. K. P., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1), 185–203. https://doi.org/10.1016/0004-3702(81)90024-2

Huang, X., Shen, C., Boix, X., & Zhao, Q. (2015). SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 262–270.

Huang, Z., & Zaidi, Q. (2022). Perceptual scale for transparency: Common fate overrides geometrical and color cues. *Journal of Vision*, *22*(6), 6. https://doi.org/10.1167/jov.22.6.6

Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2462–2470.

Jaini, P., Clark, K., & Geirhos, R. (2024). Intriguing Properties of Generative Classifiers. *The Twelfth International Conference on Learning Representations (ICLR) 2024*.

Jiang, J., Janghorbani, S., Melo, G. D., & Ahn, S. (2020). SCALOR: Generative World Models with Scalable Object Representations. *8th International Conference on Learning Representations (ICLR) 2020*.

Jiang, L., Xu, M., Liu, T., Qiao, M., & Wang, Z. (2018). DeepVS: A Deep Learning Based Video Saliency Prediction Approach. *Proceedings of the European Conference on Computer Vision (ECCV)*, 602–617.

Jiang, L., Xu, M., & Wang, Z. (2019, January 14). *Predicting Video Saliency with Object-to-Motion CNN and Two-layer Convolutional LSTM*. arXiv: 1709.06316. https://doi.org/10.1007/978-3-030-01264-9_37

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211. https://doi.org/10.3758/BF03212378

Jonschkowski, R., Stone, A., Barron, J. T., Gordon, A., Konolige, K., & Angelova, A. (2020, August). What Matters in Unsupervised Optical Flow. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 557–572). Springer International Publishing. https://doi.org/10.1007/978-3-030-58536-5_33

Joseph, K. J., Khan, S., Khan, F. S., & Balasubramanian, V. N. (2021). Towards Open World Object Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5830–5840.

Judd, T., Durand, F., & Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations. *MIT CSAIL Technical Reports* Accepted: 2012-01-13T22:30:12Z.

Kandel, E. R., Koester, J. D., Mack, S. H., & Siegelbaum, S. A. (Eds.). (2021, March 8). *Principles of Neural Science* (6th ed.). McGraw-Hill.

Karazija, L., Choudhury, S., Laina, I., Rupprecht, C., & Vedaldi, A. (2022, December). Unsupervised Multi-Object Segmentation by Predicting Probable Motion Patterns. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 2128–2141, Vol. 35). Curran Associates, Inc.

Karazija, L., Laina, I., & Rupprecht, C. (2021, December). ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* (Vol. 1).

Kellman, P. J., & Spelke, E. S. (1983). Perception of partly occluded objects in infancy. *Cognitive Psychology*, *15*(4), 483–524. https://doi.org/10.1016/0010-0285(83)90017-8

Keuper, M., Andres, B., & Brox, T. (2015). Motion Trajectory Segmentation via Minimum Cost Multicuts. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 3271–3279.

Kipf, T., Pol, E. van der, & Welling, M. (2020). Contrastive Learning of Structured World Models. *8th International Conference on Learning Representations (ICLR) 2020*.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., & Girshick, R. (2023). Segment

Anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Kirillov, A., Wu, Y., He, K., & Girshick, R. (2020). PointRend: Image Segmentation As Rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9799–9808.

Kocak, A., Erdem, E., & Erdem, A. (2022). A Gated Fusion Network for Dynamic Saliency Prediction. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(3), 995–1008. https://doi.org/10.1109/TCDS.2021.3094974

Kosiorek, A., Kim, H., Teh, Y. W., & Posner, I. (2018, December). Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 31). Curran Associates, Inc.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012, December). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25). Curran Associates, Inc.

Kruthiventi, S. S. S., Ayush, K., & Babu, R. V. (2017). DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations. *IEEE Transactions on Image Processing*, *26*(9), 4446–4456. https://doi.org/10.1109/TIP.2017.2710620

Kümmerer, M., Theis, L., & Bethge, M. (2015). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *ICLR (Workshop) 2015*.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059. https://doi.org/10.1073/pnas.1510393112

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2018, September). Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (pp. 798–814). Springer International Publishing. https://doi.org/10.1007/978-3-030-01270-0_47

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding Low- and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 4789–4798.

Lai, Q., Wang, W., Sun, H., & Shen, J. (2020). Video Saliency Prediction Using Spatiotemporal Residual Attentive Networks. *IEEE Transactions on Image Processing*, *29*, 1113–1126. https://doi.org/10.1109/TIP.2019.2936112

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lamdouar, H., Xie, W., & Zisserman, A. (2021). Segmenting Invisible Moving Objects. *32nd British Machine Vision Conference (BMVC)*.

Lamdouar, H., Yang, C., Xie, W., & Zisserman, A. (2020). Betrayed by Motion: Camouflaged Object Discovery via Motion Segmentation. *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lidén, L., & Pack, C. (1999). The role of terminators and occlusion cues in motion integration and segmentation: A neural network model. *Vision Research*, *39*(19), 3301–3320. https://doi.org/10.1016/S0042-6989(99)00055-3

Lidwell, W., Holden, K., & Butler, J. (2010, January 1). *Universal Principles of Design: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions, and Teach through Design* (2nd edition). Rockport.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015, February 21). *Microsoft COCO: Common Objects in Context*. arXiv: 1405.0312 [cs]. https://doi.org/10.48550/arXiv.1405.0312

Linardos, P., Mohedano, E., Nieto, J. J., O'Connor, N. E., Giro-i-Nieto, X., & McGuinness, K. (2019, July 16). *Simple vs complex temporal recurrences for video saliency prediction*. arXiv: 1907.01869 [cs]. https://doi.org/10.48550/arXiv.1907.01869

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023, December). Visual Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 34892–34916, Vol. 36). Curran Associates, Inc.

Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., & Kipf, T. (2020, December). Object-Centric Learning with Slot Attention. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (pp. 11525–11538, Vol. 33). Curran Associates, Inc.

Lv, W., Huang, Y., Zhang, N., Lin, R.-S., Han, M., & Zeng, D. (2024). DiffMOT: A Real-time Diffusion-based Multiple Object Tracker with Non-linear Prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19321–19330.

Mahendran, A., Thewlis, J., & Vedaldi, A. (2018). Self-Supervised Segmentation by Grouping Optical-Flow. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Mather, G., & Cavanagh, P. (1989). Motion: The long and short of it. *Spatial Vision*, *4*(2-3), 103–129. https://doi.org/10.1163/156856889X00077

Mauss, A. S., Vlasits, A., Borst, A., & Feller, M. (2017). Visual Circuits for Direction Selectivity. *Annual Review of Neuroscience*, *40*, 211–230. https://doi.org/10.1146/annurev-neuro-072116-031335

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4040–4048.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

Meister, S., Hur, J., & Roth, S. (2018). UnFlow: Unsupervised Learning of Optical Flow With a Bidirectional Census Loss. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.12276

Menze, M., Heipke, C., & Geiger, A. (2015). JOINT 3D ESTIMATION OF VEHICLES AND SCENE FLOW. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *II-3/W5*, 427–434. https://doi.org/10.5194/isprsannals-II-3-W5-427-2015

Min, K., & Corso, J. J. (2019). TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2394–2403.

Mineault, P., Zanichelli, N., Peng, J. Z., Arkhipov, A., Bingham, E., Jara-Ettinger, J., Mackevicius, E., Marblestone, A., Mattar, M., Payne, A., Sanborn, S., Schroeder, K., Tavares, Z., & Tolias, A. (2024, November 27). *NeuroAI for AI Safety*. arXiv: 2411.18526 [cs]. https://doi.org/10.48550/arXiv.2411.18526

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion. *Cognitive Computation*, *3*(1), 5–24. https://doi.org/10.1007/s12559-010-9074-z

Momennejad, I. (2022). A rubric for human-like agents and NeuroAI. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1869), 20210446. https://doi.org/10.1098/rstb.2021.0446

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., & Bowers, J. (2021). The role of Disentanglement in Generalisation. *9th International Conference on Learning Representations (ICLR) 2021*.

Needham, A. (1998). Infants' use of featural information in the segregation of stationary objects. *Infant Behavior and Development*, *21*(1), 47–76. https://doi.org/10.1016/S0163-6383(98)90054-6

Newsome, W. T., & Pare, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *Journal of Neuroscience*, *8*(6), 2201–2211. https://doi.org/10.1523/JNEUROSCI.08-06-02201.1988

Nishida, S., Kawabe, T., Sawayama, M., & Fukiage, T. (2018). Motion Perception: From Detection to Interpretation. *Annual Review of Vision Science*, *4*(1), 501–523. https://doi.org/10.1146/annurev-vision-091517-034328

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, *21*(19), 1641–1646. https://doi.org/10.1016/j.cub.2011.08.031

Oh, S. W., Lee, J.-Y., Xu, N., & Kim, S. J. (2019). Video Object Segmentation Using Space-Time Memory Networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9226–9235.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024, March 4). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs]. https://doi.org/10.48550/arXiv.2303.08774

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., … Bojanowski, P. (2024, February 2). *DINOv2: Learning Robust Visual Features without Supervision*. arXiv: 2304.07193 [cs]. https://doi.org/10.48550/arXiv.2304.07193

Petersik, J. T. (1991). Comments on Cavanagh and Mather (1989): Coming up short (and long). *Spatial Vision*, *5*(4), 291–301. https://doi.org/10.1163/156856891X00056

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, e261. https://doi.org/10.1017/S0140525X22002849

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, July). Learning Transferable Visual Models From Natural Language Supervision.

In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763, Vol. 139). PMLR.

Ramakrishnan, S. K., Wijmans, E., Kraehenbuehl, P., & Koltun, V. (2024, October 9). *Does Spatial Cognition Emerge in Frontier Models?* arXiv: 2410.06468. https://doi.org/10.48550/arXiv.2410.06468

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024, October 28). *SAM 2: Segment Anything in Images and Videos*. arXiv: 2408.00714 [cs]. https://doi.org/10.48550/arXiv.2408.00714

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788.

Ren, S., He, K., Girshick, R., & Sun, J. (2015, December). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 28). Curran Associates, Inc.

Rensink, R. A. (2000). The Dynamic Representation of Scenes. *Visual Cognition*, *7*(1-3), 17–42. https://doi.org/10.1080/135062800394667

Robert, S., Ungerleider, L. G., & Vaziri-Pashkam, M. (2023). Disentangling Object Category Representations Driven by Dynamic and Static Visual Input. *Journal of Neuroscience*, *43*(4), 621–634. https://doi.org/10.1523/JNEUROSCI.0371-22.2022

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis With Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, *39*(19), 3157–3163. https://doi.org/10.1016/S0042-6989(99)00077-2

Rushton, S. K., Bradshaw, M. F., & Warren, P. A. (2007). The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition*, *105*(1), 237–245. https://doi.org/10.1016/j.cognition.2006.09.004

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, *9*(11), 1421–1431. https://doi.org/10.1038/nn1786

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67. https://doi.org/10.1038/s41583-020-00395-8

Schott, L., Kügelgen, J. von, Träuble, F., Gehler, P., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., & Brendel, W. (2022, February 12). *Visual Representation Learning Does Not Generalize Strongly Within the Same Domain*. arXiv: 2107.08221 [cs]. https://doi.org/10.48550/arXiv.2107.08221

Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.-J., He, T., Zhang, Z., Schölkopf, B., Brox, T., & Locatello, F. (2023). Bridging the Gap to Real-World Object-Centric Learning. *The Eleventh International Conference on Learning Representations (ICLR) 2023*.

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*, 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K. C., See, S., Qin, H., Dai, J., & Li, H. (2023). FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1599–1610.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, *38*(5), 743–761. https://doi.org/10.1016/S0042-6989(97)00183-1

Simonyan, K., & Zisserman, A. (2015, April 10). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv: 1409.1556 [cs]. https://doi.org/10.48550/arXiv.1409.1556

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. https://doi.org/10.1016/0364-0213(90)90025-R

Stone, A., Maurer, D., Ayvaci, A., Angelova, A., & Jonschkowski, R. (2021). SMURF: Self-Teaching Multi-Frame Unsupervised RAFT With Full-Image Warping. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3887–3896.

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1355

Sun, D., Yang, X., Liu, M.-Y., & Kautz, J. (2018). PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8934–8943.

Sun, D., Yang, X., Liu, M.-Y., & Kautz, J. (2020). Models Matter, So Does Training: An Empirical Study of CNNs for Optical Flow Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(6), 1408–1423. https://doi.org/10.1109/TPAMI.2019.2894353

Sun, Z., Chen, Y.-J., Yang, Y.-H., & Nishida, S. (2023, December). Modeling Human Visual Motion Processing with Trainable Motion Energy Sensing and a Self-attention Network. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 24335–24348, Vol. 36). Curran Associates, Inc.

Sundaram, S., Fu, S., Muttenthaler, L., Tamir, N. Y., Chai, L., Kornblith, S., Darrell, T., & Isola, P. (2024, October 14). *When Does Perceptual Alignment Benefit Vision Representations?* arXiv: 2410.10817. https://doi.org/10.48550/arXiv.2410.10817

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014, February 19). *Intriguing properties of neural networks*. arXiv: 1312.6199 [cs]. https://doi.org/10.48550/arXiv.1312.6199

Szeliski, R. (2022, January 3). *Computer Vision: Algorithms and Applications* (2nd edition). Springer Nature Switzerland.

Teed, Z., & Deng, J. (2020, August). RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 402–419). Springer International Publishing. https://doi.org/10.1007/978-3-030-58536-5_24

Ternus, J. (1926). Experimentelle Untersuchungen über phänomenale Identität. *Psychologische Forschung*, *7*(1), 81–136. https://doi.org/10.1007/BF02424350

Tian, Z., Shen, C., & Chen, H. (2020, August). Conditional Convolutions for Instance Segmentation. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.), *Computer Vision – ECCV 2020* (pp. 282–298). Springer International Publishing. https://doi.org/10.1007/978-3-030-58452-8_17

Tian, Z., Shen, C., Chen, H., & He, T. (2019). FCOS: Fully Convolutional One-Stage Object Detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9627–9636.

Tokmakov, P., Alahari, K., & Schmid, C. (2017). Learning Motion Patterns in Videos. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3386–3394.

Tokmakov, P., Schmid, C., & Alahari, K. (2019). Learning to Segment Moving Objects. *International Journal of Computer Vision*, *127*(3), 282–301. https://doi.org/10.1007/s11263-018-1122-2

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., &

Lample, G. (2023, February 27). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs]. https://doi.org/10.48550/arXiv.2302.13971

Ullman, S. (1997). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *203*(1153), 405–426. https://doi.org/10.1098/rspb.1979.0006

Ullman, S. (1979). *The Interpretation of Visual Motion* (Third Printing, 1985). MIT Press.

Ullman, S., Harari, D., & Dorfman, N. (2012). From simple innate biases to complex visual concepts. *Proceedings of the National Academy of Sciences*, *109*(44), 18215–18220. https://doi.org/10.1073/pnas.1207690109

van Santen, J. P. H., & Sperling, G. (1984). Temporal covariance model of human motion perception. *Journal of the Optical Society of America A*, *1*(5), 451–473. https://doi.org/10.1364/JOSAA.1.000451

van Santen, J. P. H., & Sperling, G. (1985). Elaborated Reichardt detectors. *Journal of the Optical Society of America A*, *2*(2), 300–321. https://doi.org/10.1364/JOSAA.2.000300

Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., & Giro-i-Nieto, X. (2019). RVOS: End-To-End Recurrent Network for Video Object Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5277–5286.

Vig, E., Dorr, M., & Cox, D. (2014). Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2798–2805.

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Wagemans, J. (2015, August 20). *The Oxford Handbook of Perceptual Organization*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199686858.001.0001

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://doi.org/10.1037/a0029333

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., van der Helm, P. A., & van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological Bulletin*, *138*(6), 1218–1252. https://doi.org/10.1037/a0029334

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x

Wan, Z., Tang, S., Wei, J., Zhang, R., & Cao, J. (2024, October 14). *DragEntity: Trajectory Guided Video Generation using Entity and Positional Relationships*. arXiv: 2410.10751. https://doi.org/10.48550/arXiv.2410.10751

Wang, W., Shen, J., Guo, F., Cheng, M.-M., & Borji, A. (2018). Revisiting Video Saliency: A Large-Scale Benchmark and a New Model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4894–4903.

Wang, X., Girdhar, R., Yu, S. X., & Misra, I. (2023). Cut and Learn for Unsupervised Object Detection and Instance Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3124–3134.

Wang, X., Yang, J., & Darrell, T. (2024, December). Segment Anything without Supervision. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in Neural Information Processing Systems* (pp. 138731–138755, Vol. 37). Curran Associates, Inc.

Watson, A. B., & Ahumada, A. J. (1985). Model of human visual-motion sensing. *Journal of the Optical Society of America A*, *2*(2), 322–342. https://doi.org/10.1364/JOSAA.2.000322

Weis, M. A., Chitta, K., Sharma, Y., Brendel, W., Bethge, M., Geiger, A., & Ecker, A. S. (2021). Benchmarking Unsupervised Object Representations for Video Sequences. *Journal of Machine Learning Research*, *22*(183), 1–61.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. https://doi.org/10.1038/nn0602-858

Wertheimer, M. (1912). Experimentelle Studien über das Sehen von Bewegung. (F. Schumann & J. Rich. Ewald, Eds.). *Zeitschrift für Psychologie*, *61*, 161–265.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II. *Psychologische Forschung*, *4*(1), 301–350. https://doi.org/10.1007/BF00410640

Wichmann, F. A., & Geirhos, R. (2023). Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, *9*, 501–524. https://doi.org/10.1146/annurev-vision-120522-031739

Wixted, J. T., & Serences, J. T. (Eds.). (2018). *Steven's Handbook of Experimental Psychology and Cognitive Neuroscience, Volume 2: Sensation, Perception, & Attention*. John Wiley & Sons, Ltd.

Wolfe, J. M. (2000, January 1). Visual Attention. In K. K. De Valois (Ed.), *Seeing* (pp. 335–386). Academic Press. https://doi.org/10.1016/B978-012443760-9/50010-6

Wu, S., He, X., Lu, H., & Yuille, A. L. (2010). A unified model of short-range and long-range motion perception. *Advances in Neural Information Processing Systems*, *23*.

Xie, J., Xie, W., & Zisserman, A. (2022, December). Segmenting Moving Objects via an Object-Centric Layered Representation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (pp. 28023–28036, Vol. 35). Curran Associates, Inc.

Xie, J., Yang, C., Xie, W., & Zisserman, A. (2024, April 18). *Moving Object Segmentation: All You Need Is SAM (and Flow)*. arXiv: 2404.12389. https://doi.org/10.48550/arXiv.2404.12389

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., & Tao, D. (2022). GMFlow: Learning Optical Flow via Global Matching. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8121–8130.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yang, Y., Loquercio, A., Scaramuzza, D., & Soatto, S. (2019). Unsupervised Moving Object Detection via Contextual Information Separation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 879–888.

Yang, Y.-H., Fukiage, T., Sun, Z., & Nishida, S. (2023). Psychophysical measurement of perceived motion flow of naturalistic scenes. *iScience*, *26*(12). https://doi.org/10.1016/j.isci.2023.108307

Yarbus, A. L. (1967). *Eye Movements and Vision*. Springer. https://doi.org/10.1007/978-1-4899-5379-7

Yuan, J., Chen, T., Li, B., & Xue, X. (2023). Compositional Scene Representation Learning via Reconstruction: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 11540–11560. https://doi.org/10.1109/TPAMI.2023.3286184

Zadaianchuk, A., Seitzer, M., & Martius, G. (2023, December). Object-Centric Learning for Real-World Videos by Predicting Temporal Feature Similarities. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 61514–61545, Vol. 36). Curran Associates, Inc.

Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., DiCarlo, J., Ganguli, S., Hawkins, J., Koerding, K., Koulakov, A., LeCun, Y., Lillicrap, T., Marblestone, A., Olshausen, B., … Tsao, D. (2023, February 22). *Toward Next-Generation*

Artificial Intelligence: Catalyzing the NeuroAI Revolution. arXiv: 2210.08340 [cs]. https://doi.org/10.48550/arXiv.2210.08340

Zarei Eskikand, P., Kameneva, T., Ibbotson, M. R., Burkitt, A. N., & Grayden, D. B. (2016). A Possible Role for End-Stopped V1 Neurons in the Perception of Motion: A Computational Model. *PLOS ONE*, *11*(10), e0164813. https://doi.org/10.1371/journal.pone.0164813

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2023). Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(4), 4396–4415. https://doi.org/10.1109/TPAMI.2022.3195549

Appendix: Original publications

The following peer-reviewed publications are included in this thesis:

- Matthias Tangemann, Matthias Kümmerer, Thomas S.A. Wallis, Matthias Bethge (2020). *Measuring the importance of temporal features in video saliency.* Computer Vision – ECCV 2020. Lecture Notes in Computer Science, vol 12373. Springer, Cham.
- Matthias Tangemann, Steffen Schneider, Julius von Kügelgen, Francesco Locatello, Peter Gehler, Thomas Brox, Matthias Kümmerer, Matthias Bethge, Bernhard Schölkopf (2023). *Unsupervised object learning via common fate.* Proceedings of the Second Conference on Causal Learning and Reasoning, PMLR 213:281-327.
- Matthias Tangemann, Matthias Kümmerer, Matthias Bethge (2024). *Object segmentation from common fate: Motion energy processing enables human-like zero-shot generalization to random dot stimuli.* Advances in Neural Information Processing Systems 37 (NeurIPS 2024).

More detailed information about the contributions of each author are listed in the chapters discussing the respective publication.

Measuring the Importance of Temporal Features in Video Saliency

Matthias Tangemann[1(✉)], Matthias Kümmerer[1], Thomas S.A. Wallis[1,2], and Matthias Bethge[1,2]

[1] University of Tübingen, Tübingen, Germany
{matthias.tangemann,matthias.kuemmerer,tom.wallis,matthias}@bethgelab.org
[2] Amazon Research, Tübingen, Germany

Abstract. Where people look when watching videos is believed to be heavily influenced by temporal patterns. In this work, we test this assumption by quantifying to which extent gaze on recent video saliency benchmarks can be predicted by a static baseline model. On the recent LEDOV dataset, we find that at least 75% of the explainable information as defined by a gold standard model can be explained using static features. Our baseline model "DeepGaze MR" even outperforms state-of-the-art video saliency models, despite deliberately ignoring all temporal patterns. Visual inspection of our static baseline's failure cases shows that clear temporal effects on human gaze placement exist, but are both rare in the dataset and not captured by any of the recent video saliency models. To focus the development of video saliency models on better capturing temporal effects we construct a meta-dataset consisting of those examples requiring temporal information.

Keywords: Gaze prediction · Saliency · Video · Temporal modelling · Model evaluation

1 Introduction

The human visual system processes information from the environment selectively. Several attention mechanisms limit the amount of information to be processed and thus enable efficient perception of the world (e.g., [9]). The most obvious form of attention is the shifting of gaze, which orients the high-resolution fovea towards areas of interest.

Modelling those gaze shifts is an important topic in computer vision. Predictive models of human gaze have the potential to advance our understanding of human visual attention, for example by aiding the development of hypotheses that can be tested with human subjects [7]. Besides their scientific usefulness, such models have various technical applications. They can be used for graphic

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58604-1_40) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12373, pp. 667–684, 2020.
https://doi.org/10.1007/978-3-030-58604-1_40

design [6], automated cropping, video compression [11] or other computer vision tasks (e.g., [48]).

Great progress has been made recently in predicting where people look in still images. With the use of pre-trained models the performance improved from 1/3 to more than 80% of explainable information explained (e.g., [25,27]). Since the human visual system developed in a dynamic environment, there is growing interest to also model human gaze on videos. Previous studies revealed that motion patterns are an important factor attracting visual attention [8,16,39]. All recent video gaze models therefore are based on temporal modeling components such as recurrent units or spatiotemporal convolutions to capture those dynamic patterns.

To which degree those temporal patterns influence human gaze on natural videos and to which degree the recent performance improvements in video gaze prediction can be attributed to capturing these effects, however, has not been evaluated thoroughly so far. With our work we aim at filling this gap, by providing a method to measure the influence of temporal patterns on human gaze. We construct a static baseline model that by design cannot capture temporal effects and compare its performance to a gold standard model estimating the total information in the ground truth gaze data. The performance difference to the gold standard then represents an upper bound to the influence of temporal effects on the respective dataset. Furthermore, by looking at the largest failure cases of our static baseline, we can identify situations in the dataset where human gaze is driven by temporal patterns. Evaluating gaze prediction models on those situations then lets us draw conclusions about the capabilities of models to predict temporal effects.

Applying this method to the recent LEDOV dataset [20] and state-of-the-art video gaze models we arrive at the following conclusions: (1) Human gaze placement on the videos contained in the LEDOV dataset is largely driven by spatial appearance. (2) Clearly identifiable temporal effects on human visual attention exist, but occur rarely in the videos considered. (3) We need to construct suitable video data sets to enable learning based models to capture temporal effects. Indeed, we show that all other recent video gaze models with the capacity for temporal modelling fail in the same situations as our restricted model.

We explicitly note that the main contribution of our work are above findings and the proposed evaluation method that we need to come to those findings, but not the static baseline model that is required for our analysis. Interestingly though, our baseline model outperforms state-of-the-art video gaze prediction models on the LEDOV and DIEM [34] datasets—despite deliberately ignoring all temporal information.

To enable other researches to apply our proposed evaluation method more easily, we collect a meta-benchmark from existing datasets that contains the situations requiring temporal information revealed by our analysis. The performance of new models on this meta-benchmark indicates how much an improved predictive performance can be attributed to better handling of temporal effects. We will make this meta-benchmark as well as our pre-trained baseline model publicly available.

2 Related Work

Substantial progress has been made on the task of gaze prediction for free viewing of images. While the influential model by Itti and Koch [18], inspired by Treisman and Gelade's feature integration theory [45], was devised to explain effects observed in visual search originally, it also achieved first successes in predicting where people look. Since then, more than 50 models have been proposed predicting probable gaze locations based on image content (for a recent comparison see, e.g., [12]). As in other areas of computer vision, the advent of deep learning gave rise to models greatly improving state-of-the-art performance [15,24,27,35,46]. DeepGaze II [27], the current state of the art model on the MIT Saliency Benchmark [5], captures 81% of the explainable information gain on that dataset (explainable information gain is an information-theoretic analogue of explainable variance, see [25] for details).

In contrast, gaze prediction for videos only recently attracted more attention. Several datasets and models have been developed, but neither a widely accepted benchmark nor an estimate of the amount of explainable information in those datasets exist. This makes an evaluation of the state of the field difficult.

Recently, two video gaze datasets have been introduced that are large enough to train deep neural network based models: LEDOV [20] and DHF1K [47]. More recently, Wang et al. also provided gaze recordings for video segmentation datasets [48]. The gaze recordings provided by Mathe and Sminichescu [32] for the Hollywood and UCF-Sports dataset are large enough for deep learning based approaches, but most of the subjects have not been recorded in the free-viewing setting. Several small datasets exists that provide high-quality recordings (e.g., DIEM [34], for an overview see [20]).

Starting with an extension of the Itti and Koch model to videos [16,17], several models predicting gaze specifically for videos have been proposed [10,13,14,30,38,40,41,51–53]. The performance of video gaze models has been greatly improved with the advent of deep learning. Bazzani et al. [3] trained a recurrent neural network based on features extracted from a spatiotemporal DNN predicting gaze using a mixture of Gaussians. The models by Wang et al. [47] and Wu et al. [50] pair convolutional LSTM units with an attention mechanism. Bak et al. [2] proposed a two-stream network using optical flow in parallel to RGB features. This two-stream approach has also been combined with convolutional LSTM units by [19,20] and with convolutional GRU and an attention mechanism by [28]. Linardos et al. [31] proposed a model based on an exponential moving average of frame-wise features. Very recently, [33] and [43] proposed spatio-temporal encoder-decoder networks for video gaze prediction.

3 Methods

The main objective of our work is to evaluate the influence of temporal patterns on human gaze. To that end, we propose a baseline model that cannot learn temporal patterns by design but predicts human gaze on videos solely relying on static appearance. This baseline model is then compared to a gold standard model as an estimate of the total information in the ground truth gaze data. The performance difference between those models represents an upper bound of the influence of temporal patterns on human gaze placement.

3.1 Center Bias

The center bias is an important lower baseline. It is obtained by blurring and normalizing a histogram of all gaze positions in the training set. As humans tend to look at the center of images [44] and videos are usually recorded such that interesting objects are in the middle of videos, there is a clear bias in the gaze data towards the center of the videos. The center bias therefore represents a prior distribution of gaze position independent of visual content. Predicting this spatial prior for every frame yields a lower baseline, comparable to the chance level performance in classification tasks.

The center bias is much stronger in the beginning of each video due to the subjects fixating the center of the screen before each trial. As described later, we ignore this effect by not evaluating on the first 15 frames and confirmed experimentally that a stationary center bias models the remaining data well. Furthermore, we optimized the blur size using a grid search.

3.2 Gold Standard Model

The maximal performance that gaze prediction models can achieve is limited by the consistency of the subjects and varies from frame to frame. We use a gold standard model [49] to measure the inter-subject variability of the gaze positions. The model predicts where each subject looked given the ground truth information from all other subjects on the same frame. This is done by blurring the gaze positions of all but one subject and performing leave-one-out cross validation. Moreover, the prediction of the gold standard model is mixed with a uniform distribution to handle outliers. The gold standard therefore predicts subjects to look where other subjects look with a high probability, and to randomly look anywhere on the image with a small probability defined by the mixing coefficient. The optimal blur size of the gaussian filter and the mixing weight of the uniform distribution are determined using a grid search.

A high gold standard performance indicates very consistent gaze locations across all subjects and vice versa. Therefore, the gold standard model yields an estimate of the maximal performance that can be achieved for every frame. All reported gold standard performances refer to the leave-one-subject-out performance.

3.3 Static Baseline Model

Our proposed evaluation method requires a static baseline model that cannot handle temporal effects by design. Initial experiments revealed that DeepGaze II [27], the current state-of-the-art model for images, achieves a very competitive performance when simply applied to videos frame-by-frame (see Sect. 4). However, this instantaneous model ignores delays due to the required processing in the human brain. This suggests a way to improve the DeepGaze II architecture for videos by averaging deep features over multiple recent video frames. Based on this approach, we propose a space-time separable variant of DeepGaze II using a temporal box filter as static baseline model (see Fig. 1), which we call *DeepGaze Mean Readout (DeepGaze MR)*.

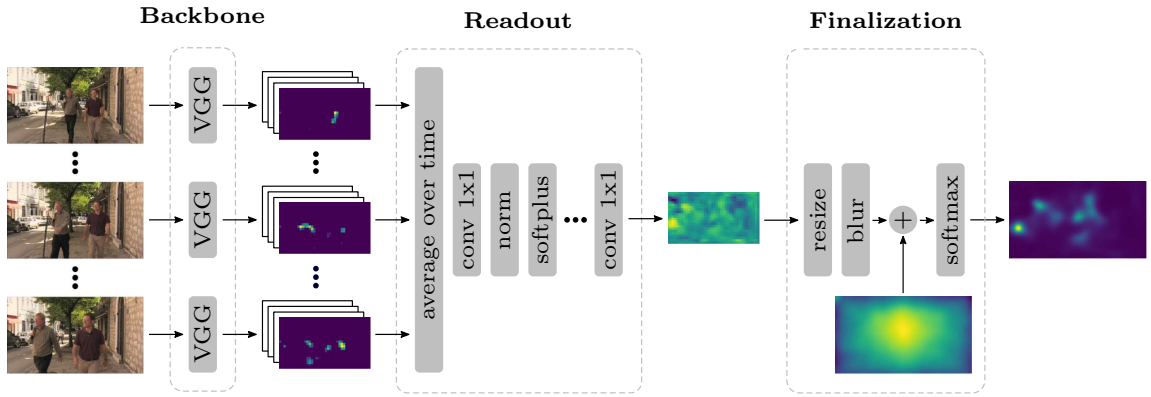

Fig. 1. Architecture of our static baseline model "DeepGaze MR": A feature representation is extracted from individual frames in a fixed size window using the VGG-19 network. A non-linear readout network transforms this representation into a priority map by first averaging the feature channels over time, and then applying a series of 1×1 convolutions. The resulting map is then resized, blurred, weighted by the center bias, and normalized to obtain the final prediction.

Input to our model is a fixed length window of consecutive frames, which is used to predict the gaze distribution on the last frame ("target frame") in this window. We use a window length of 16 frames, which was the optimal value found using a grid search (see supplement for details).

Backbone. Our model applies the VGG-19 network pretrained on Imagenet [42] to every frame individually and extracts the representation from the last convolutional layer (`conv5_4`) after the nonlinearity. We keep the parameters of the backbone fixed to prevent overfitting.

Readout. A non-linear readout network is used to transform the feature representation into a priority map of probable gaze locations. The readout network first averages the feature representation over time. A series of 1×1 pixel convolutions is then used to non-linearly combine the feature channels to the priority map. Layer Normalization [1] is used after all but the last convolutional layer

to stabilize training. As non-linearity we use the softplus function, which is a smooth approximation of the commonly used ReLU and avoided units zeroing out early in training. We use three convolutional layers with 32, 32, and one channel, respectively. This optimal instantiation of the readout network has been found using a random search (see supplement for details).

Finalization. Finally, the output of the readout network is turned into the predicted probability distribution: First, the priority map is resized to the resolution of the input. It is then smoothed using a Gaussian with learnable standard deviation per x and y dimension. The logarithm of the center bias density from the training set is added to the map using a learnable weight, acting as a spatial prior. Finally, a softmax is applied to obtain the predicted spatial probability distribution of gaze locations.

Training. Our model is trained using maximum-likelihood learning (Kümmerer et al. [26] suggest that this allows for best metric scores in all classic saliency metrics). Thus, the loss function is the average log-density at gaze locations for each frame. We use the Adam optimizer [23] with a learning rate of 0.01, which is decreased by a factor of ten after one and five epochs. In each epoch, only one random target frame per video is used for training. Experiments confirmed, that this training scheme is sufficient for our model to converge.

Since our model averages features over time, it is by design not able to represent temporal patterns such as movements, or appearing and disappearing objects.

4 Experiments

In this section, we evaluate our baseline models described above on recent video gaze datasets. We then analyse the baseline predictions in comparison to state-of-the-art video gaze models to better understand the importance of temporal effects in video saliency.

The evaluation of gaze prediction models comes with challenges: different evaluation protocols and metrics typically lead to inconsistent model rankings. Building on recent work to better understand this evaluation process, we first describe and motivate the model evaluation approach used in this work.

4.1 Metrics

A large number of metrics exist that are used to evaluate gaze predictions (for a review see [4]). As typically used, these metrics give rise to inconsistent model rankings. Kümmerer et al. [26] proposed to adapt a probabilistic setting, i.e., to formulate models so that they predict spatial probability distributions, train them for log-likelihood and differentiate between predictions and derived saliency maps. In this way, consistent model ratings can be achieved.

We adopt this setting in our work, and use information gain (average log-likelihood per fixation compared to the center bias, [25]) as our primary metric.

To enable a comparison to models that did not use a probabilistic approach, we additionally evaluate the AUC [22], NSS [36], CC [21], KLDiv [29,37] and SIM [22] metrics to judge the performance of our model relative to state-of-the-art. To obtain an overall score for a model, the metrics are applied to the prediction for every frame individually, and then averaged first over frames and then over videos.

4.2 Datasets

The main dataset for this work is the LEDOV dataset [19]. It contains 538 short videos (11s on average) with eye tracking data of 32 subjects. The authors removed smooth pursuits and saccades and artificially stabilized fixations during their preprocessing, so this dataset does not allow to investigate the precise dynamics of individual gaze trajectories. However, the dataset covers the common factors driving human gaze placement sufficiently well to develop and compare models. All videos have been rescaled to 640×360px and resampled to 30 Hz. Models from other groups are evaluated using the resolutions and frame rates that the respective models have been trained on.

For additional analyses we are using the DIEM dataset [34] (84 videos, 66 subjects on average, mean duration 95.2 s). The videos have been padded to match the viewing conditions reported in the paper and rescaled to 640×480px.

The DHF1K dataset [47] is comparable in scope to LEDOV, but contains artifacts in the provided gaze maps. As those artifacts affect the model scores and make it impossible to properly evaluate the gold standard model, we excluded this dataset from our analysis. In the supplemental information we provide more details on this issue together with overall performance results which suggest that our conclusions are also valid for DHF1K.

For all datasets, the subject had to fixate the center of the screen before each trial. We do not evaluate models on the first 15 frames to ignore the centered gaze due to the experimental paradigm.

4.3 Performance Results

In Table 1, we show the performances of our baselines and other recent gaze models on LEDOV. Despite deliberately ignoring all temporal effects, DeepGaze MR performs very well and explains as much as 75% of the explainable information (as a comparison, the state-of-the-art for images on MIT1003 is 81%). Moreover, DeepGaze MR performs substantially better than DeepGaze II which confirms the effectiveness of our proposed adaptations. Interestingly, in AUC our model matches the gold standard performance, which might be due to the fact that AUC saturates very quickly. Furthermore, the AUC metric might suffer from the leave-one-subject-out cross validation applied in the gold standard.

We further compare the performances of our baselines to recent video gaze prediction models: The DeepVS model [19,20] allows the most direct comparison, as it was trained on the LEDOV dataset as well. ACLNet [47], SalEMA [31], TASED-Net [33] and STRA-Net [28] are recent video gaze models developed

on the DHF1K dataset [47]. For all models, we used the published weights and adapted size and frame rate of the input videos to match the samples encountered in the respective model training. As the results in Table 1 show, DeepGaze MR clearly outperforms all evaluated previous state-of-the-art models on the LEDOV dataset across all metrics, despite being designed as a static baseline model.

Table 1. Performance comparison of recent gaze prediction models on the LEDOV dataset. The information gain can only be evaluated for models that predict a spatial probability distribution. All models have been applied using the published weights. TASED-NET, SalEMA, ACLNet and STRA-Net have been trained on the DHF1K dataset, DeepGaze II on SALICON and MIT1003.

LEDOV val

Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.833	0.157	3.521	1.546	0.062
TASED-Net [33]	-	-	0.887	0.647	3.214	3.498	0.496
STRA-Net [28]	-	-	0.890	0.610	2.315	3.324	0.460
SalEMA [31]	-	-	0.890	0.596	2.573	3.331	0.466
ACLNet [47]	-	-	0.892	0.587	1.905	3.156	0.430
DeepVS [19]	-	-	0.894	0.397	2.445	3.098	0.210
DeepGaze II [27]	1.216	62.8	0.908	0.588	1.259	3.368	0.434
DeepGaze MR	**1.445**	**74.6**	**0.917**	**0.665**	**1.105**	**3.857**	**0.498**
Gold standard	1.961	100	0.917	-	-	4.992	-

LEDOV test

Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.844	0.142	3.689	1.585	0.057
SalEMA [31]	-	-	0.897	0.590	2.377	3.152	0.465
TASED-Net [33]	-	-	0.897	0.650	2.965	3.361	0.505
ACLNet [47]	-	-	0.898	0.573	1.667	2.922	0.435
STRA-Net [28]	-	-	0.899	0.597	2.024	3.130	0.466
DeepVS [19]	-	-	0.903	0.394	2.398	3.081	0.218
DeepGaze II [27]	1.117	61.0	0.909	0.606	1.195	3.403	0.447
DeepGaze MR	**1.367**	**75.5**	**0.920**	**0.667**	**1.035**	**3.657**	**0.506**
Gold standard	1.810	100	0.920	-	-	4.676	-

Additionally, we evaluated the models on the DIEM dataset. The size of the dataset is rather small (84 videos), therefore we did not train but only evaluate the models on this dataset. As the results in Table 2 show, our model performs clearly better than all other video saliency methods on this dataset except TASED-Net [33]. Interestingly, the original DeepGaze II model for images performs even better than the variant adapted to videos.

The performances on DIEM are worse than those on LEDOV for two reasons. First, this dataset is much harder as the videos in this dataset contain much more temporal activity. Second, the domain gap to LEDOV is rather large, as DIEM contains cuts and many objects not present in LEDOV. The good performance of DeepGaze II on this dataset could therefore be explained by the much broader range of objects it has seen during training. Moreover, DeepGaze II is applied purely frame-by-frame, so it probably copes better with the many cuts in DIEM.

4.4 Analyzing Temporal Effects

In the following, we try to better understand the influence of temporal information on gaze placement. As motivated earlier, we use the information gain difference of the gold standard and DeepGaze MR as an estimator of the information that is not captured by our model. Since the model cannot learn temporal patterns by design, temporal effects on human gaze placement should result in large differences to the gold standard.

In Fig. 2 we plot the distribution of those differences grouped by video. The median remaining information is close to 0bit for roughly half of the videos in the validation set. This indicates that our static baseline model successfully predicts gaze positions on a large number of frames. However, the results also clearly show two kinds of failure cases: (1) There are some videos for which the average performance gap to the gold standard is large. For the first three videos in the plot, the median difference is almost 2bit. (2) For other videos there is a large number of outlier frames whose performance gap is much greater than for most of the other remaining frames in the video. As our model is not able to exploit temporal structure by design, they should include cases in which temporal patterns affect human gaze placement.

We analyze the found failure cases in more detail by visualizing them in Figs. 3 and 4. We plot the NSS scores of the models over time (bottom) and visualize the model log predictions at interesting frames (top, frame position indicated by dashed lines in the NSS plot). As SalEMA averages features and thus cannot handle temporal effects by design, we don't consider it in this case study. The figures reveal three common factors that strongly influence where people look and are difficult for all models:

Interactions between objects occur in several of the videos. Here, most subjects look at the interaction point, not at the objects themselves. This is clearly observable in Fig. 4b, when the child is feeding the giraffe or in Fig. 3a when the robot is grabbing objects.

Suddenly appearing objects have a very strong ability to attract human attention as well. As can be seen in Fig. 4a the shifting of the gaze to the appearing text is very consistent across all subjects. We assume that this effect can be observed with suddenly appearing objects in general, but cannot verify this hypothesis properly due to the small number of samples. A related effect is the appearing of the two persons due to the camera motion in Fig. 4d. They also clearly attract attention, however much less than the sudden appearing of the text in Fig. 4a.

Table 2. Performance comparison of recent gaze prediction models on the full DIEM dataset. Due to the small number of videos, none of the models has been trained on this dataset.

	DIEM						
Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.892	0.436	1.511	2.053	0.341
DeepVS [19]	-	-	0.853	0.424	2.070	2.096	0.309
SalEMA [31]	-	-	0.911	0.576	1.743	2.987	0.465
STRA-Net [28]	-	-	0.914	0.595	1.975	3.069	0.477
TASED-Net [33]	-	-	0.914	**0.621**	2.098	**3.194**	**0.493**
ACLNet [47]	-	-	0.914	0.558	1.468	2.826	0.428
DeepGaze MR	0.660	43.1	0.920	0.602	1.091	3.116	0.471
DeepGaze II [27]	**0.674**	**44.0**	**0.926**	0.619	**1.058**	2.898	0.477
Gold standard	1.531	100	0.940	-	-	4.659	-

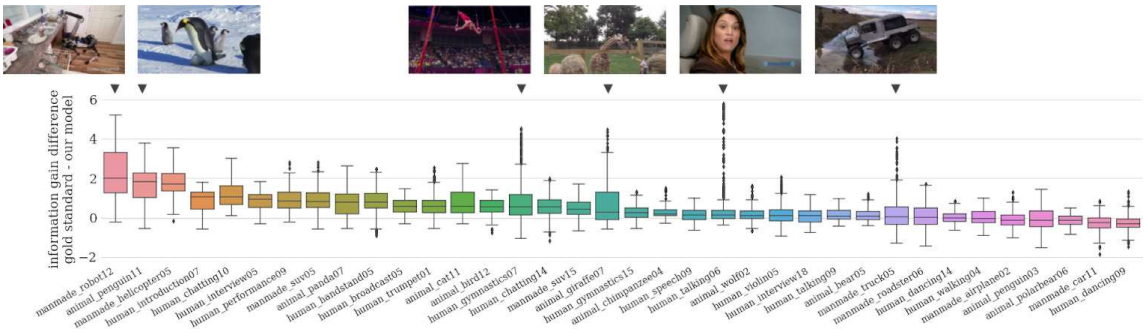

Fig. 2. Distribution of the unexplained information across frames in the LEDOV validation set (x-axis shows distinct videos). The remaining explainable information is estimated by the difference of our model to the gold standard in bit using the information gain metric. The videos marked are the largest failure cases of our model.

Movements of objects also clearly have the potential to change which parts of a scene are observed. In Fig. 4c, none of the subjects looks at the gymnast's arms or hands, but all are looking at his torso that is moving in the respective scene. This stands in contrast to most cases in which humans appear, where subjects tend to look at people's hands or faces. Also global camera movements seem to be able to shift people's gaze towards the side of the direction of the movement, as indicated in Fig. 4d. However the effect in this example is small and entangled with the appearing persons. A closer investigation would be necessary to address this effect.

The temporal effects described are compiled from the qualitative analysis of our model's largest failure cases. As their number is small, the list given is most likely not exhaustive. Moreover, it is not possible to reliably draw any general conclusions about the relative strengths of those effects. However, the cases presented clearly reveal the existence of such temporal effects and show

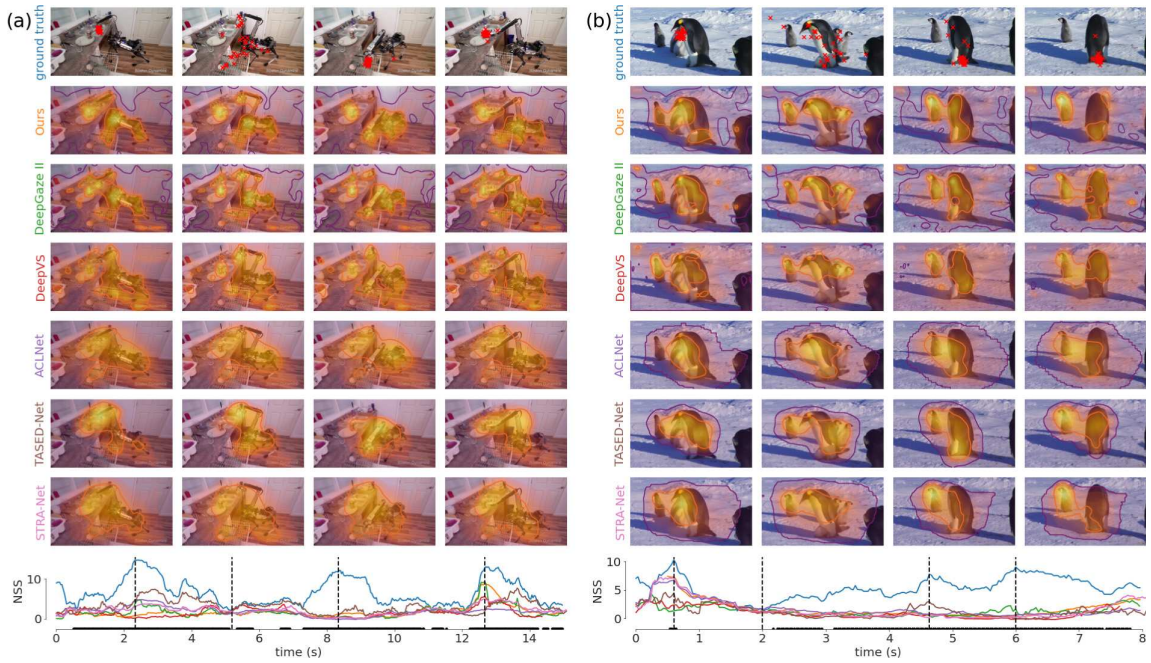

Fig. 3. Failure cases with a high average difference to the gold standard: **(a)** Most of the subjects look at the robot's hand while it puts a glass into a dishwasher. The models however distribute their prediction over the whole robot. **(b)** After roughly two seconds a small penguin becomes visible under the big penguin in the foreground, shifting the gaze of most subjects to the small penguin for the rest of the video. Markers on the x-axis of the NSS plots indicate frames that are part of our proposed meta-benchmark (see Sect. 4.5).

that they are not captured by all recent video gaze models that should have the capacity to model them.

4.5 Evaluating Temporal Modelling

The detailed analysis of the failure cases in the previous section showed that none of the considered models was able to correctly predict cases in which temporal information influences where people look. As our proposed method requires training and evaluating two baseline models, the hurdle to apply it is quite high. To facilitate applying our method to new models, we propose a principled new meta-benchmark consisting of those hard cases.

Our meta-dataset contains all frames of videos where our static baseline's information gain is at least 1bit worse than the gold standard (indicated by markers on the x-axis of the NSS plots in Figs. 3 and 4). We propose to run the models on the full videos, but only average the performances over the frames included in our meta-dataset. This evaluation scheme discards roughly 80% of the frames in LEDOV and 65% of the frames in DIEM. As our model cannot learn temporal effects by design, gaze on the discarded frames can be explained by spatial features. The performance on the remaining frames reflects the ability of models to handle cases in which temporal information is necessary much better than existing benchmarks.

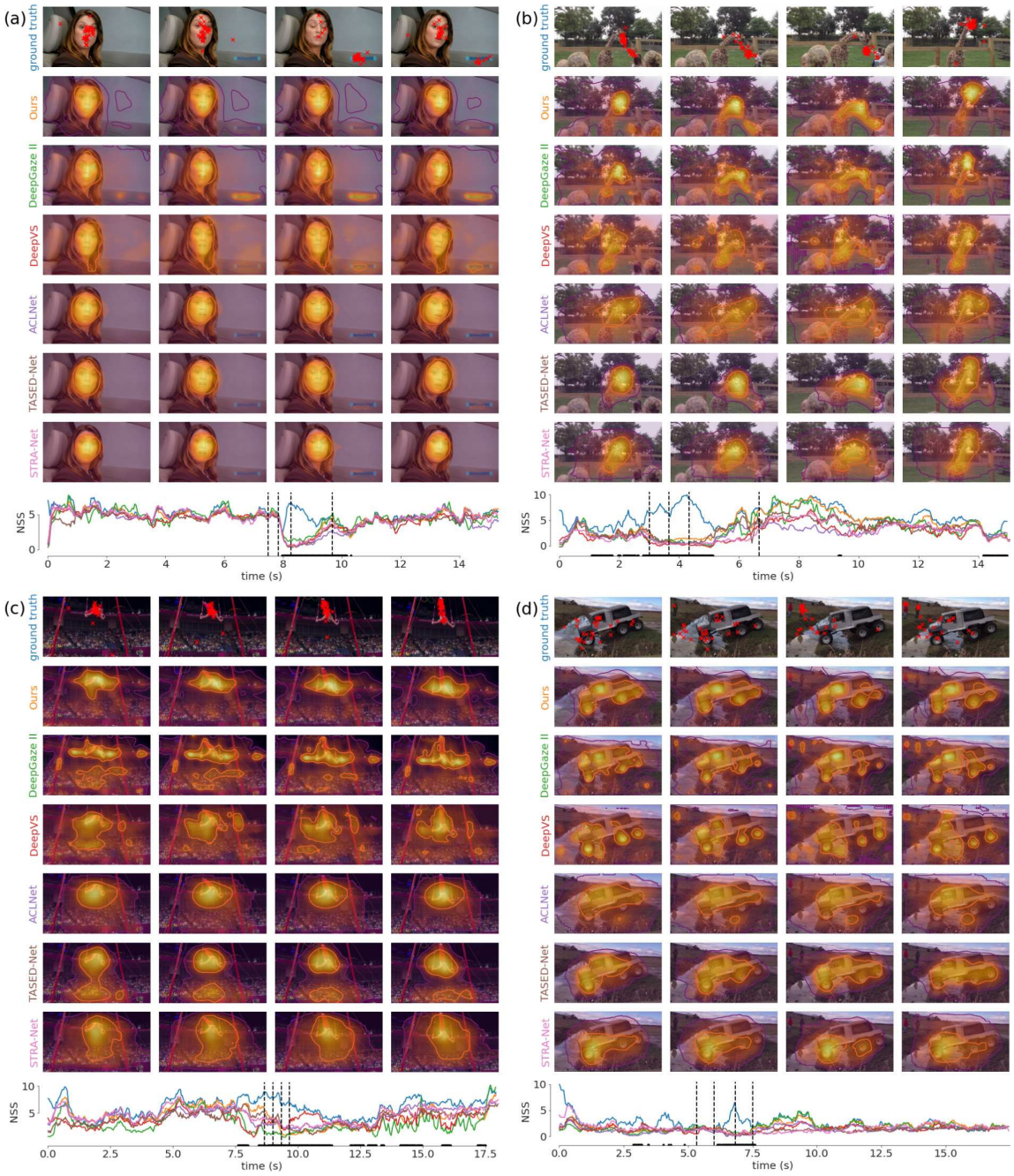

Fig. 4. Failure cases due to localized events: **(a)** A text suddenly appearing draws almost all attention for a short time, whereas the models predict people to mainly look at the person talking. **(b)** When a child is feeding a giraffe, the subjects' attention focuses at their interaction point and not at the giraffe's head that is looked at during the remainder of the video. **(c)** Gaze concentrates on the gymnast's torso during a swinging exercise, whereas other body parts are much less looked at. **(d)** Two persons enter the scene due to the movement of the camera, which temporarily attracts the attention of most of the subjects.

Table 3. Performance of state-of-the-art models on our proposed meta-benchmark, which discards frames in which the information gain of our model is more than 1bit less as the gold standard. As our model cannot exploit temporal patterns, the reported performances reflect the ability to handle cases in which temporal information is needed to predict where people look much better.

	Meta-Benchmark: LEDOV & DIEM						
Model	IG	%	AUC	CC	KLdiv	NSS	SIM
Center bias	0	0	0.853	0.274	2.580	1.679	0.195
DeepVS [19]	-	-	0.854	0.337	2.599	2.152	0.225
SalEMA [31]	-	-	0.887	0.477	2.584	2.596	0.394
STRA-Net [28]	-	-	0.889	0.497	2.681	2.658	0.39
ACLNet [47]	-	-	0.891	0.483	2.044	2.579	0.377
TASED-Net [33]	-	-	0.893	**0.583**	2.995	**2.855**	**0.430**
DeepGaze MR	0.528	24.2	0.898	0.454	1.568	2.458	0.365
DeepGaze II [27]	**0.787**	**36.1**	**0.908**	0.507	**1.420**	2.693	0.389
Gold Standard	2.182	100.0	0.948	-	-	5.093	-

In Table 3 we report the model performances on this meta-benchmark derived from LEDOV and DIEM. As indicated by our previous analysis, all models considered in this work perform poorly. As this benchmark was derived from failure cases of our model, the performance reduction of our model is disproportionally large. When using DeepGaze II as a baseline model, our model performs much better in this meta-benchmark (see supplement for details).

5 Discussion

Human gaze on dynamic stimuli such as videos is hypothesized to be strongly driven by temporal patterns in the stimuli, e.g., temporal popup and motion [8,16]. In this work, we measured the importance of temporal features in video saliency. To that end, we developed and analysed DeepGaze MR, a static baseline model predicting gaze positions on the LEDOV dataset, and compared it's performance to a gold standard model. DeepGaze MR is adapted from the successful DeepGaze II model for still images and is not able to learn temporal patterns by design. Nevertheless, our model outperforms previous state of the art with a large margin on the LEDOV dataset and captures 75% of the explainable information gain.

When we analyzed failure cases of our model, we found clear temporal effects that drove the subject's gaze such as sudden appearances and movements and, to a certain degree, also interactions. We found that the gold standard performance and therefore the consistency among subjects is very high in those cases. This confirms the hypothesis that temporal patterns are an important factor influencing human gaze placement.

Given this importance of temporal effects, we would expect a good video saliency model to predict human gaze in those cases well. While our model wasn't able to capture those effects by design, we found that all other models we tested consistently failed to capture those effects either. In particular, this is the case also for models like DeepVS, ACLNet, STRA-Net and TASED-Net that have explicitly been designed to capture temporal patterns.

We argue that the main reason for this shortcoming is a deficiency of the datasets used to train video saliency models. Temporal patterns in the videos can influence gaze placement in ways that are highly consistent over subjects (Fig. 4, see also [8,16]). However, these effects turn out to be rare compared to the influence of spatial patterns such as faces on gaze placement. We suppose that they are so rare that current state-of-the-art models do not benefit from investing modelling capacity into modelling them. This difficulty for learning-based approaches to handle rare, but important, events correctly is a general problem relevant for many fields. In autonomous driving, for example, it is crucial to handle rare events correctly, e.g. when children running onto the street.

Several aspects can contribute to tackling this issue: the model architecture, the loss function and the datasets.

Adding general temporal modelling components, as done by previous works on video saliency, has shown to be ineffective to learn temporal effects. However, our study reveals distinct temporal effects on human gaze. Models might benefit from adding modules that are explicitly designed to detect effects that we know to be relevant, such as appearing objects.

To evaluate models predicting gaze on videos, image-based metrics are typically applied per frame and averaged. As a result, some of the failure cases seen above do not substantially affect the overall model performance as those effects tend to be short compared to the whole video. This is opposed to our subjective impression of the clear failure of the model on those samples. A loss function that penalizes such failures more visibly would align the benchmark results better with human judgement.

We see the most fundamental need for improvement in the datasets. Obviously, one could explicitly collect and add cases of relevant temporal patterns to the training datasets. In particular, it would be possible to have multiple validation datasets tailored towards effects that might be considered relevant, such as appearing objects, motion and interactions. In this way, one can quantitatively judge how well new models incorporate effects that researchers consider relevant for understanding behaviour, but that are rare in the usual training datasets.

Finally, we introduced a meta-benchmark derived from existing datasets that allows to quantify the ability of models to handle those temporal effects much better: Instead of averaging performances over all frames, we only consider frames in which the information gain of our model is more than 1bit smaller than the gold standard. As our model cannot learn temporal patterns, only frames are discarded in which spatial information is sufficient. The low performance of existing models on this meta-dataset confirms our previous analysis. We will make a list of the frames we considered in this study available. In the future,

our proposed benchmark could be improved by considering more datasets and by improving our spatial baseline model.

Acknowledgements. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): Germany's Excellence Strategy – EXC 2064/1 – 390727645 and SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 3, project number: 276693517. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Matthias Tangemann.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer Normalization. arXiv:1607.06450 [cs, stat], July 2016
2. Bak, C., Kocak, A., Erdem, E., Erdem, A.: Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Trans. Multimed **20**(7), 1688–1698 (2018). https://doi.org/10.1109/TMM.2017.2777665
3. Bazzani, L., Larochelle, H., Torresani, L.: Recurrent mixture density network for spatiotemporal visual attention. In: ICLR 2017 (2017)
4. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 185–207 (2013). https://doi.org/10.1109/TPAMI.2012.89
5. Bylinskii, Z., et al.: MIT saliency benchmark. http://saliency.mit.edu/
6. Bylinskii, Z., et al.: Learning visual importance for graphic designs and data visualizations. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology, UIST 2017, pp. 57–69. ACM, New York (2017). https://doi.org/10.1145/3126594.3126653
7. Cichy, R.M., Kaiser, D.: Deep neural networks as scientific models. Trends Cogn. Sci. **23**(4), 305–317 (2019). https://doi.org/10.1016/j.tics.2019.01.009
8. Dorr, M., Martinetz, T., Gegenfurtner, K.R., Barth, E.: Variability of eye movements when viewing dynamic natural scenes. J. Vis. **10**(10), 28–28 (2010). https://doi.org/10.1167/10.10.28
9. Eysenck, M.W., Keane, M.T.: Cognitive Psychology: A Student's Handbook, vol. 6. Psychology Press, London (2010)
10. Fang, Y., Lin, W., Chen, Z., Tsai, C.M., Lin, C.W.: A video saliency detection model in compressed domain. IEEE Trans. Circuits Syst. Video Technol. **24**(1), 27–38 (2014). https://doi.org/10.1109/TCSVT.2013.2273613
11. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans. Image Process. **19**(1), 185–198 (2010). https://doi.org/10.1109/TIP.2009.2030969
12. He, S., Tavakoli, H.R., Borji, A., Mi, Y., Pugeault, N.: Understanding and visualizing deep visual saliency models. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10206–10215, June 2019
13. Hossein Khatoonabadi, S., Vasconcelos, N., Bajic, I.V., Shan, Y.: How many bits does it take for a stimulus to be salient? In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5501–5510, June 2015

14. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 21, pp. 681–688. Curran Associates, Inc. (2009)
15. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks. In: The IEEE International Conference on Computer Vision (ICCV), pp. 262–270, December 2015
16. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cogn. **12**(6), 1093–1123 (2005). https://doi.org/10.1080/13506280444000661
17. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194–203 (2001). https://doi.org/10.1038/35058500
18. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998). https://doi.org/10.1109/34.730558
19. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: DeepVS: a deep learning based video saliency prediction approach. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 625–642. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_37
20. Jiang, L., Xu, M., Wang, Z.: Predicting Video Saliency with Object-to-Motion CNN and Two-layer Convolutional LSTM. arXiv:1709.06316 [cs], September 2017
21. Jost, T., Ouerhani, N., von Wartburg, R., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. Comput. Vis. Image Underst. **100**(1–2), 107–123 (2005). https://doi.org/10.1016/j.cviu.2004.10.009
22. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. MIT Technical report, January 2012
23. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR 2015, May 2015
24. Kümmerer, M., Theis, L., Bethge, M.: Deep gaze I: boosting saliency prediction with feature maps trained on ImageNet. In: ICLR Workshops 2015, May 2015
25. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Information-theoretic model comparison unifies saliency metrics. Proc. Natl. Acad. Sci. **112**(52), 16054–16059 (2015). https://doi.org/10.1073/pnas.1510393112
26. Kümmerer, M., Wallis, T.S.A., Bethge, M.: Saliency benchmarking made easy: separating models, maps and metrics. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11220, pp. 798–814. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01270-0_47
27. Kümmerer, M., Wallis, T.S., Gatys, L.A., Bethge, M.: Understanding low- and high-level contributions to fixation prediction. In: The IEEE International Conference on Computer Vision (ICCV), pp. 4799–4808, October 2017
28. Lai, Q., Wang, W., Sun, H., Shen, J.: Video saliency prediction using spatiotemporal residual attentive networks. IEEE Trans. Image Process. **29**, 1113–1126 (2020). https://doi.org/10.1109/TIP.2019.2936112
29. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behav. Res. Methods **45**(1), 251–266 (2013). https://doi.org/10.3758/s13428-012-0226-9
30. Leborán, V., García-Díaz, A., Fdez-Vidal, X.R., Pardo, X.M.: Dynamic whitening saliency. IEEE Trans. Pattern Anal. Mach. Intell. **39**(5), 893–907 (2017). https://doi.org/10.1109/TPAMI.2016.2567391

31. Linardos, P., Mohedano, E., Nieto, J.J., O'Connor, N.E., Giro-i-Nieto, X., McGuinness, K.: Simple vs complex temporal recurrences for video saliency prediction. In: British Machine Vision Conference (BMVC), September 2019
32. Mathe, S., Sminchisescu, C.: Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(7), 1408–1424 (2015). https://doi.org/10.1109/TPAMI.2014.2366154
33. Min, K., Corso, J.J.: TASED-Net: temporally-aggregating spatial encoder-decoder network for video saliency detection. In: The IEEE International Conference on Computer Vision (ICCV), pp. 2394–2403, October 2019
34. Mital, P.K., Smith, T.J., Hill, R.L., Henderson, J.M.: Clustering of gaze during dynamic scene viewing is predicted by motion. Cogn. Comput. **3**(1), 5–24 (2011). https://doi.org/10.1007/s12559-010-9074-z
35. Pan, J., et al.: SalGAN: visual saliency prediction with generative adversarial networks. arXiv:1701.01081 [cs], January 2017
36. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision. Res. **45**(18), 2397–2416 (2005). https://doi.org/10.1016/j.visres.2005.03.019
37. Rajashekar, U., Cormack, L.K., Bovik, A.C.: Point-of-gaze analysis reveals visual search strategies. In: Human Vision and Electronic Imaging IX, vol. 5292, pp. 296–306. International Society for Optics and Photonics, June 2004. https://doi.org/10.1117/12.537118
38. Ren, Z., Gao, S., Chia, L.T., Rajan, D.: Regularized feature reconstruction for spatio-temporal saliency detection. IEEE Trans. Image Process. **22**(8), 3120–3132 (2013). https://doi.org/10.1109/TIP.2013.2259837
39. Rosenholtz, R.: A simple saliency model predicts a number of motion popout phenomena. Vision. Res. **39**(19), 3157–3163 (1999). https://doi.org/10.1016/S0042-6989(99)00077-2
40. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1147–1154, June 2013
41. Seo, H.J., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. J. Vis. **9**(12), 15–15 (2009). https://doi.org/10.1167/9.12.15
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR 2015, May 2015
43. Sun, Z., Wang, X., Zhang, Q., Jiang, J.: Real-time video saliency prediction via 3d residual convolutional neural network. IEEE Access **7**, 147743–147754 (2019). https://doi.org/10.1109/ACCESS.2019.2946479
44. Tatler, B.W.: The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. J. Vis. **7**(14), 4–4 (2007). https://doi.org/10.1167/7.14.4
45. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980). https://doi.org/10.1016/0010-0285(80)90005-5
46. Vig, E., Dorr, M., Cox, D.: Large-scale optimization of hierarchical features for saliency prediction in natural images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2798–2805 (2014)
47. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: a large-scale benchmark and a new model. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4894–4903, June 2018
48. Wang, W., et al.: Learning unsupervised video object segmentation through visual attention. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3064–3074, June 2019

49. Wilming, N., Betz, T., Kietzmann, T.C., König, P.: Measures and limits of models of fixation selection. PLoS ONE **6**(9), e24038 (2011). https://doi.org/10.1371/journal.pone.0024038
50. Wu, X., Wu, Z., Zhang, J., Ju, L., Wang, S.: SalSAC: a video saliency prediction model with shuffled attentions and correlation-based ConvLSTM. In: Thirty-Fourth AAAI Conference on Artificial Intelligence. AAAI Press, February 2020
51. Zhang, L., Tong, M.H., Cottrell, G.W.: SUNDAy: saliency using natural statistics for dynamic analysis of scenes. In: Proceedings of the 31st Annual Meeting of the Cognitive Science Society, pp. 2944–2949. AAAI Press, Cambridge (2009)
52. Zhong, S.h., Liu, Y., Ren, F., Zhang, J., Ren, T.: Video saliency detection via dynamic consistent spatio-temporal attention modelling. In: Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press, July 2013
53. Zhou, F., Bing Kang, S., Cohen, M.F.: Time-mapping using space-time saliency. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3358–3365, June 2014

Supplemental Material for: Measuring the Importance of Temporal Features in Video Saliency

Matthias Tangemann[1][0000−0001−9734−8692], Matthias Kümmerer[1][0000−0001−9644−4703], Thomas S.A. Wallis[1,2][0000−0001−7431−4852], and Matthias Bethge[1,2]

[1] University of Tübingen, Tübingen, Germany
[2] Amazon Research, Tübingen, Germany
{matthias.tangemann,matthias.kuemmerer,tom.wallis,matthias}@bethgelab.org

1 DHF1K Dataset

The gaze maps of the DHF1K dataset [6] contain artifacts in the gaze maps that make it impossible to properly evaluate the gold standard model and most likely affect model scores. Therefore, as stated in the main paper, we did not evaluate models on DHF1K. Here we provide more details on this issue.

For the DHF1K dataset, gaze positions have been collected from 17 subjects and provided as binary gaze maps for every frame. In Figure 1 we plot a histogram of the number of gaze positions per frame. The histogram clearly shows that substantially more positions than subjects are given per gaze map for all frames. On average, the number of gaze positions is ten times higher than the number of subjects. The example gaze maps in Figure 2 show that the subject's gaze positions are represented as irregular clusters of multiple pixels and large, grid-like structures in the map.

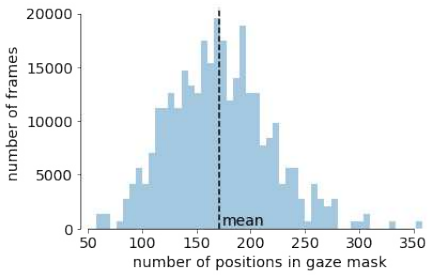

Fig. 1: Histogram of the number of positions in the binary gaze maps provided by the DHF1K dataset (17 subjects).

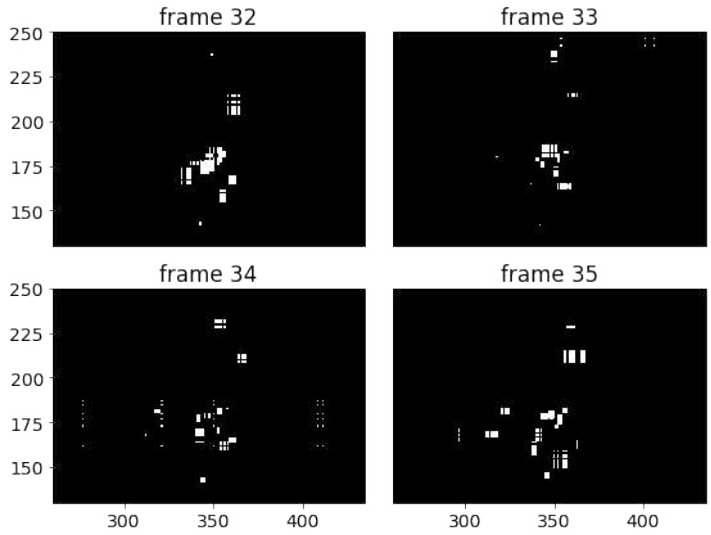

Fig. 2: Example gaze maps from sample "0601" of the DHF1K dataset. The maps represent gaze positions as irregular clusters of several pixels and contain large, grid-like structures.

The gold standard model is based on leave-one-out cross validation across subjects. The gaze maps provided with the DHF1K dataset however don't allow to determine the gaze locations for the individual subjects which makes it impossible to properly evaluate the gold standard model. Furthermore we expect those artifacts to affect the metrics used to evaluate gaze prediction models: depending on how many pixels are contained in any of the pixel clusters (which have very diverse sizes), that cluster will contribute more or less to the loss on this frame. For those reasons, we do not use the DHF1K dataset in our work.

Nevertheless, we evaluated the performance of our proposed model on the DHF1K validation set. As the results in table 1 show, DeepGaze MR performs better than many video saliency models. We did not adapt any hyperparameters for this dataset, so it is likely that the performance of our method could be improved further. The best performing model SalEMA [4] is based on an exponentially moving average, so similarly to our baseline model it cannot model temporal effects by design either. Summing up, those results suggest that temporal effects are of minor importance also for DHF1K.

2 Architecture Search

The architecture of DeepGaze MR described in the main paper has some important hyperparameters: We determined the number of input frames using a grid

	DHF1K		
Model	IG	AUC	NSS
Center bias	0	0.853	1.674
DeepVS [1]	-	0.854	1.067
DeepGaze II [2]	0.238	0.881	1.833
ACLNet [6]	-	0.893	2.412
DeepGaze MR	**0.702**	0.897	2.587
TASED-Net [5]	-	0.901	2.822
SalEMA [4]	-	**0.905**	**2.849**
Gold Standard	-	-	-

Table 1: Performance of state-of-the-art models on DHF1K. Due to the artifacts in the provided gaze maps the gold standard performance cannot be evaluated.

search and the depth and number of channels in the readout network using a random search. In the following, we present the respective results in more detail.

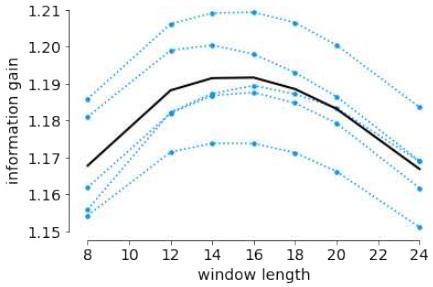

Fig. 3: Performance of a linear readout of VGG features averaged over time for different window lengths. Each blue line represents a single iteration, for which the same seed has been used for all window lengths. The black line represents the average performance per window length.

To find the optimal **window length** we used a linear instantiation of our model (i.e., only one convolutional layer in the readout network). We trained this model as the model described in the main paper, except that we used a learning rate of 0.001 which was decreased by a factor of 10 after 4 epochs. For the grid search we trained the model using 9 different window lengths from 8 to 24 frames and repeated the grid search for 5 different seeds. For all window lengths, the first 32 frames have been ignored for each video.

According to the results show in Figure 3, the optimal window length is 16 frames. However, the parameters seems to be not too sensitive as window lengths of 12–18 frames yielded a very similar performance.

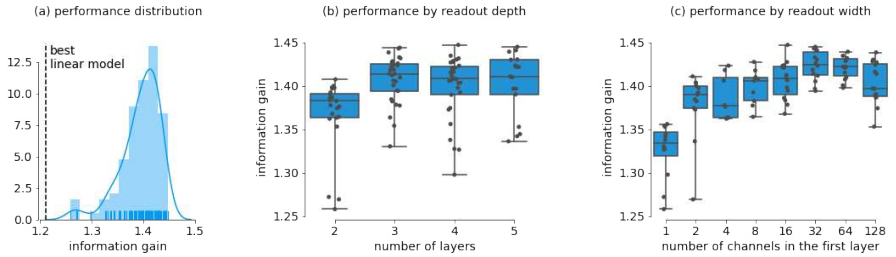

Fig. 4: Results of the random search for the optimal configuration of the readout network. (a) The distribution of performances achieved. The dashed line indicates the performance of the best linear model. (b) The performance by the number of layers in the readout network. (c) The performance by the number of channels in the first layer of the readout network.

Using the optimal window length of 16 frames, we then performed a random search to find the optimal **architecture of the readout network**. We trained 100 models having two to five layers and [1, 2, 4, ..., 128] channels in each layer.

The results are summarized in Figure 4. As the distribution of model performances in Figure 4a shows, most models clearly outperform the best linear model from the previous experiment. So a non-linear readout network is clearly needed for a good model. However, several different configurations of the readout network achieved a similar and good performance. So the detailed configuration of the readout network seems to be of minor importance.

Nevertheless, the random search revealed some trends. In Figure 4b, we plot the model performance of the networks by the number of convolutional layers used. The results indicate that the readout network should have a depth of at least 3 layers. A readout network consisting of only two layers is clearly too shallow and typically performs worse than deeper models.

Finally, we analyze the model performance depending on the number of channels in the first convolutional layer. Typically, this layer contains the majority of the readout network's parameters and therefore has a large impact on the model's capacity. As the results Figure 4c show, the number of channels in the first layer has to be high enough. Having only one layer doesn't allow the model to learn feature interactions and is clearly outperformed by the readout networks with higher capacity. Having 32 channels in the first layer seems to be optimal. Interestingly, we didn't see any signs of overfitting even for the biggest readout networks. So moderately sized readout networks appear sufficient to capture all available information for our architecture.

One of the simplest models reaching a top performance has 3 layers with 32, 32 and 1 channel, respectively. This is the architecture presented in the main paper.

Meta-Benchmark: LEDOV & DIEM							
Model	IG	%	AUC	CC	KLDiv	NSS	SIM
Center bias	0	0	0.869	0.258	2.607	1.918	0.190
DeepVS [1]	-	-	0.855	0.327	2.580	2.115	0.218
SalEMA [4]	-	-	0.890	0.470	2.454	2.613	0.389
ACLNet [6]	-	-	0.893	0.473	1.916	2.547	0.369
STRA-Net [3]	-	-	0.896	0.498	2.345	2.699	0.397
TASED-Net [5]	-	-	0.903	**0.567**	2.625	**3.078**	**0.448**
DeepGaze II [2]	0.326	14.8	0.903	0.445	1.545	2.041	0.354
DeepGaze MR	**0.799**	**36.2**	**0.913**	0.543	**1.325**	3.069	0.419
Gold Standard	2.207	100	0.952	-	-	5.490	-

Table 2: Performance of state-of-the-art models on a variant of our proposed meta-benchmark, using DeepGaze II as a baseline instead of DeepGaze MR. Comparing to the results of the original meta-benchmark, DeepGaze MR achieves much better results.

3 DeepGaze II as Baseline for the Meta-Benchmark

In our main work, we proposed a meta-benchmark consisting of those frames for which the information gain of the new DeepGaze MR model is more than 1bit worse than the gold standard model. By using the DeepGaze MR model as a baseline for defining the new benchmark, its results are disproportionally worse than that of other models.

As a comparison, we report the results of the benchmark using DeepGaze II as baseline model in Table 2. This way, our model performs similar to DeepGaze II on the original meta-benchmark whereas DeepGaze II now performs substantially worse. The results of the remaining models consistently improved compared to the original meta-benchmark. This indicates, that the benchmark variant defined using DeepGaze II is easier than the benchmark proposed in the paper.

References

1. Jiang, L., Xu, M., Liu, T., Qiao, M., Wang, Z.: DeepVS: A Deep Learning Based Video Saliency Prediction Approach. In: The European Conference on Computer Vision (ECCV). pp. 602–617 (Sep 2018)
2. Kümmerer, M., Wallis, T.S., Gatys, L.A., Bethge, M.: Understanding Low- and High-Level Contributions to Fixation Prediction. In: The IEEE International Conference on Computer Vision (ICCV). pp. 4799–4808 (Oct 2017)
3. Lai, Q., Wang, W., Sun, H., Shen, J.: Video Saliency Prediction using Spatiotemporal Residual Attentive Networks. IEEE Transactions on Image Processing **29**, 1113–1126 (2020). https://doi.org/10.1109/TIP.2019.2936112
4. Linardos, P., Mohedano, E., Nieto, J.J., O'Connor, N.E., Giro-i-Nieto, X., McGuinness, K.: Simple vs complex temporal recurrences for video saliency prediction. In: British Machine Vision Conference (BMVC) (Sep 2019)
5. Min, K., Corso, J.J.: TASED-Net: Temporally-Aggregating Spatial Encoder-Decoder Network for Video Saliency Detection. In: The IEEE International Conference on Computer Vision (ICCV). pp. 2394–2403 (Oct 2019)
6. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting Video Saliency: A Large-scale Benchmark and a New Model. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4894–4903 (Jun 2018)

Unsupervised Object Learning via Common Fate

Matthias Tangemann [1,2] MATTHIAS.TANGEMANN@BETHGELAB.ORG
Steffen Schneider [1,2,3] STEFFEN.SCHNEIDER@BETHGELAB.ORG
Julius von Kügelgen [4,5] JVK@TUEBINGEN.MPG.DE
Francesco Locatello [6] LOCATELF@AMAZON.COM
Peter Gehler [6] PGEHLER@AMAZON.COM
Thomas Brox [6] BROX@AMAZON.DE
Matthias Kümmerer [1,2,*] MATTHIAS.KUEMMERER@BETHGELAB.ORG
Matthias Bethge [1,2,*] MATTHIAS.BETHGE@UNI-TUEBINGEN.DE
Bernhard Schölkopf [6,*] BERNHARD@AMAZON.COM

[1] *Tübingen AI Center* [2] *University of Tübingen* [3] *EPFL* [4] *Max Planck Institute for Intelligent Systems, Tübingen* [5] *University of Cambridge* [6] *Amazon* [*] *Joint senior authors*

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Learning generative object models from unlabelled videos is a long standing problem relevant for causal scene modeling. We decompose this task into three easier subtasks, and provide candidate solutions for each of them. Inspired by the Common Fate Principle of Gestalt Psychology, we first extract (noisy) masks of moving objects via unsupervised motion segmentation. Second, generative models are trained on the masks of the background and the moving objects, respectively. Third, background and foreground models are combined in a conditional "dead leaves" scene model to sample novel scene configurations where occlusions and depth layering arise naturally. To evaluate the individual stages, we introduce the FISHBOWL dataset positioned between complex real-world scenes and common object-centric benchmarks of simplistic objects. We show that our approach learns generative models that generalize beyond occlusions present in the input videos and represents scenes in a modular fashion, allowing generation of plausible scenes outside the training distribution by permitting, for instance, object numbers or densities not observed during training.
Code: https://github.com/mtangemann/common_fate_object_learning

Keywords: object learning, scene modeling, scene generation, causal modeling, causal representation learning, generative modeling, common fate

1. Introduction

Machine learning excels if sufficient training data is available that is representative of the task at hand. In recent years, this *i.i.d.* data paradigm has been shown not only to apply for pattern recognition problems, but also for generative modeling (Goodfellow et al., 2014). In practice, the amount of data required to reach a given level of performance will depend on the dimensionality of the data. The generation of high-dimensional images thus either requires huge amounts of data (Karras et al., 2020) or methods that exploit prior information, for instance on multi-scale structure or compositionality (Razavi et al., 2019).

Imagine we would like to automatically generate realistic images of yearbook group photos. A "brute force" approach would be to collect a massive dataset and train a large GAN (Goodfellow

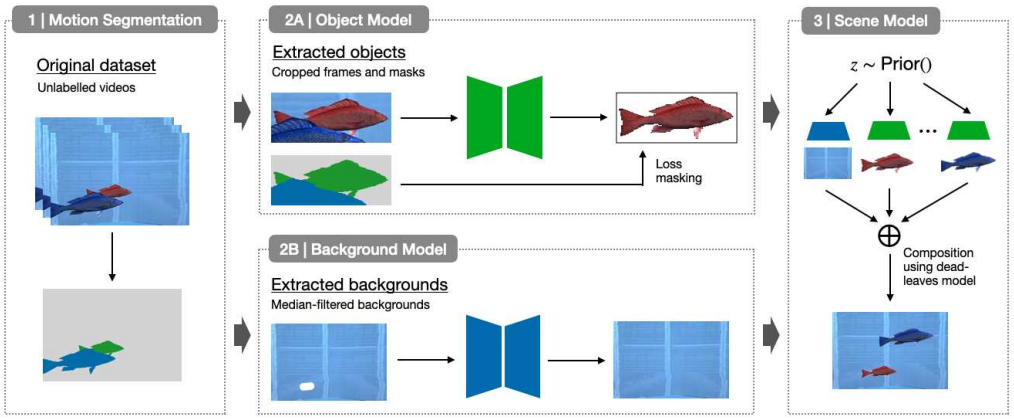

Figure 1: We propose a multi-stage approach for learning object-centric generative scene models: (1) Motion segmentation detects moving objects in the input videos. The predicted (noisy) segmentation masks are used to (2A) extract object crops for training a generative object model and (2B) extract backgrounds for training a generative background model. (3) A scene model combines the object and background models to sample novel scenes, permitting interventions on variables such as the background and the number and locations of fish.

et al., 2014), hoping that the model will not only learn typical backgrounds, but also the shape of individual humans (or human faces), and arrange them into a group. A more modular approach, in contrast, would be to learn *object models* (e.g., for faces, or humans), and learn in which positions and arrangements they appear, as well as typical backgrounds. This approach would be more data-efficient: each training image would contain multiple humans, and we would thus effectively have more data for the object learning task. In addition, the sub-task would be lower-dimensional than the original task. Finally, if we left the i.i.d. setting (by, say, having a second task with different group sizes), some of the sub-models could be re-used and the modular approach would thus lend itself more readily to knowledge transfer and "out-of-distribution" (o.o.d.) generalization.

Object-centric approaches aim to capture this compositionality and have been improved over the past few years (e.g., Locatello et al. (2020); Engelcke et al. (2021)). However, these models tend to be difficult to train and do not yet scale well to visually more complex scenes. In addition, the commonly employed end-to-end learning approaches make it difficult to dissect the causes of these difficulties and what principles may be crucial to facilitate unsupervised object learning.

In human vision, the *Principle of Common Fate* of Gestalt Psychology (Wertheimer, 2012), which posits that elements that are moving together tend to be perceived as one, has been shown to play an important role for object learning (Spelke, 1990). In our work, we show that this principle can be successfully used also for machine vision as part of a multi-stage object learning approach (Fig. 1): First, we use unsupervised motion segmentation to obtain a candidate segmentation of a video frame. Second, we train generative object and background models on this segmentation. While the regions obtained by the motion segmentation are caused by objects moving in 3D, only *visible* parts can be segmented. To learn the actual objects (i.e., the *causes*), a crucial task for the object model is learning to generalize beyond the occlusions present in its input data. To measure success, we provide a dataset including object ground truth. As the last stage, we show that the learned object and background models can be combined into an interventional world model (Schölkopf et al., 2021) that allows generating *manipulated* novel scenes. Thus, in contrast to existing object-

centric models trained end-to-end, we decompose object learning into evaluable subproblems and test the potential of exploiting object motion for building scalable object-centric models that allow for causally meaningful interventions in the scene generation process.

Summing up, the present work makes the following contributions:

- We provide the novel FISHBOWL dataset, positioned between simplistic toy scenarios and real world data, providing ground truth information for evaluating causal scene models.
- We show that the *Common Fate Principle* can be succesfully used for object learning by proposing a multi-stage object learning approach based on this principle.
- We demonstrate that the generative object and background models learned in this way can be combined into flexible scene models allowing for controlled generation of novel images.

The dataset with rendering code and models including training code is available at https://github.com/mtangemann/common_fate_object_learning.

2. Related Work

Modular scene modeling. The idea to individually represent objects in a scene is not new. One approach, motivated by the analysis-by-synthesis paradigm from cognitive science (Bever and Poeppel, 2010), assumes a detailed specification of the generative process and infers a scene representation by trying to invert this process (Kulkarni et al., 2015; Wu et al., 2017; Jampani et al., 2015). Many methods instead aim to also learn the generative process in an unsupervised way, see Greff et al. (2020) for a recent survey. Several models use a recurrent approach to sequentially decompose a given scene into objects (Eslami et al., 2016; Stelzner et al., 2019; Kosiorek et al., 2018; Gregor et al., 2015; Mnih et al., 2014; Yuan et al., 2019; Engelcke et al., 2020; Weis et al., 2020; von Kügelgen et al., 2020; Burgess et al., 2019), or directly learn a partial ordering (Heess et al., 2011; Le Roux et al., 2011). This sequential approach has also been extended with spatially-parallel components by Dittadi and Winther (2019); Jiang et al. (2019); Lin et al. (2020b); Chen et al. (2020). Other methods infer all object representations in parallel (Greff et al., 2017; van Steenkiste et al., 2018), with subsequent iterative refinement (Greff et al., 2019; Veerapaneni et al., 2019; Locatello et al., 2020; Nanbo et al., 2020). Whereas most of the above models are trained using a reconstruction objective—usually in a variational framework (Kingma and Welling, 2014; Rezende et al., 2014)—several works have also extended GANs (Goodfellow et al., 2014) to generate scenes in a modular way (Yang et al., 2017; Turkoglu et al., 2019; Nguyen-Phuoc et al., 2020; Ehrhardt et al., 2020; Niemeyer and Geiger, 2021). Those approaches typically use additional supervision such as ground-truth segmentation or additional views, with Ehrhardt et al. (2020); Niemeyer and Geiger (2021) being notable exceptions. While most methods can decompose a given scene into its constituent objects, only few are fully-generative in the sense that they can generate novel scenes (Lin et al., 2020a; Ehrhardt et al., 2020; Engelcke et al., 2020; von Kügelgen et al., 2020; Engelcke et al., 2021; Niemeyer and Geiger, 2021; Dittadi and Winther, 2019).

Our approach differs from previous works in the following three key aspects. First, previous approaches typically train a full scene model in an end-to-end fashion and include architectural biases that lead to the models decomposing scenes into objects. While elegant in principle, those methods have not been shown to scale to more realistic datasets yet. Using a multi-stage approach, as in the present work, enables re-use of existing computer vision methods (such as unsupervised motion segmentation) for well-studied sub-tasks and therefore scales more easily to visually complex

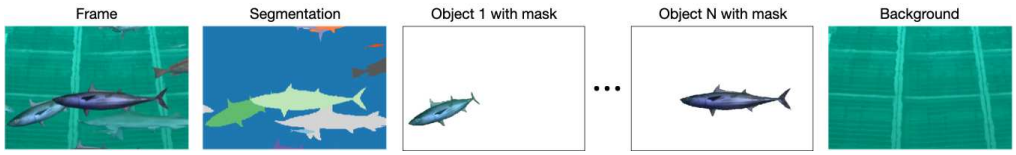

Figure 2: The FISHBOWL dataset: each video contains 128 frames with ground truth segmentation; renderings and masks (without occlusion) of every fish and background are provided in the validation and test sets.

scenes. Second, while some existing methods make use of temporal information from videos (Kipf et al., 2022; Lin et al., 2020a; Crawford and Pineau, 2020; Kosiorek et al., 2018; Weis et al., 2020), they do not explicitely use motion signals to discover (i.e., segment) objects or require weak supervision (Kipf et al., 2022). Inspired by the development of the human visual system (Spelke, 1990), we instead explicitly include this segmentation cue in our approach. Third, most existing fully-generative approaches use a spatial mixture model to compose objects into a scene (Ehrhardt et al., 2020; Engelcke et al., 2020, 2021). While this simplifies training, it clearly does not match the true underlying scene generation process. In this work, we instead follow the dead leaves model-approach of von Kügelgen et al. (2020) and scale it to more complex scenes.

Motion Segmentation. We require an unsupervised motion segmentation method that is able to segment multiple object instances. For this, we build on a line of work that tracks points with optical flow, and then performs clustering in the space of the resulting point trajectories (Brox and Malik, 2010; Ochs et al., 2014; Ochs and Brox, 2012; Keuper et al., 2015). Motion segmentation methods that require supervision (Dave et al., 2019; Xie et al., 2019; Tokmakov et al., 2017a,b) or only perform binary motion segmentation (Yang et al., 2021, 2019; Ranjan et al., 2019) are not applicable in our unsupervised setting.

Learning from motion. In the present work, we propose exploiting motion information to decompose a scene into objects, and to learn generative object and scene models. Motion information is believed to be an important cue for the development of the human visual system (Spelke, 1990) and has also been used as a training signal for computer vision (Bao et al., 2022; Chen et al., 2022; Meunier et al., 2022; Pathak et al., 2017; Mahendran et al., 2018a,b). While similar in spirit, these works however do not address learning of generative object and scene models.

3. The Fishbowl Dataset

Several video datasets have been used for object-centric represention learning before (e.g., Greff et al. (2022); Weis et al. (2020); Yi et al. (2019); Ehrhardt et al. (2020); Kosiorek et al. (2018)). The ground truth object masks and appearances provided by those datasets, however, only cover the visible parts of the scene. In order to evaluate the capabilities of the object model to infer and represent the full objects even in the presence of occlusions, we propose the novel FISHBOWL dataset positioned between complex real world data and simplistic toy datasets. This dataset consist of 20,000 training and 1,000 validation and test videos recorded from a publicly available WebGL demo of an aquarium,[1] each with a resolution of 480×320px and 128 frames. We adapted the rendering to obtain ground truth segmentations of the scene and the ground truth unoccluded background and objects (Fig. 2). More details regarding the recording setup can be found in the supplement.

1. http://webglsamples.org/aquarium/aquarium.html, 3-clause BSD license

4. A multi-stage approach for unsupervised scene modelling

We model an image $\mathbf{x}$ as a composition of a background $(\mathbf{m}_0 = \mathbb{1}, \mathbf{x}_0)$ and an ordered list of objects $(\mathbf{m}_i, \mathbf{x}_i)$, each represented as a binary mask $\mathbf{m}_i$ and appearance $\mathbf{x}_i$. The background and objects are composed into a scene using a simple "dead leaves" model, i.e., the value of each pixel (u, v) is determined by the foremost object covering that pixel. We propose a multi-stage approach (Fig. 1) for learning generative models representing objects, backgrounds and scenes in this fashion.

STAGE 1: MOTION SEGMENTATION—OBTAINING CANDIDATE OBJECTS FROM VIDEOS

As a first step, we use unsupervised motion segmentation to obtain candidate segmentations of the input videos. We build on the minimum cost multicut method by Keuper et al. (2015), which tracks a subset of the pixels through the video using optical flow and then, inspired by the *Common Fate Principle* mentioned earlier, clusters the trajectories based on pairwise motion affinities. To obtain background masks for training the background model, it is not necessary to differentiate between multiple object instances. Hence, we use an ensemble of different background-foreground segmentation models. More details are provided in the appendix.

STAGE 2A: OBJECT MODEL—LEARNING TO GENERATE UNOCCLUDED, MASKED OBJECTS

Object extraction. We use the bounding boxes of the candidate segmentation to extract object crops from the original videos and rescale them to a common size of 128×64px. We filter out degenerate masks by ignoring all masks with an area smaller than 64 pixels and only consider bounding boxes with a minimum distance of 16px to the frame boundary. Accordingly, we extract the candidate *segmentation masks* $\mathbf{m}_0, \ldots, \mathbf{m}_K$ for each crop. For notational convenience, we take $\mathbf{m}_0$ and $\mathbf{m}_1$ to correspond to the background and the object of interest (i.e., the object whose bounding box was used to create the crop), respectively, so that $\mathbf{m}_k$ with $k \geq 2$ correspond to masks of other objects.

Task. We use the segmented object crops for training a β-VAE-based generative object model (Higgins et al., 2017). Input to the model is the object crop without the segmentation, output is the reconstructed object appearance including the binary object mask. We train the model with the standard β-VAE loss with an adapted reconstruction term including both the appearance and the mask. For an input batch, let $\mathbf{c}$ and $\mathbf{m}_{0:K}$ be the ground truth crops with candidate segmentations, and $\hat{\mathbf{c}}$ and $\hat{\mathbf{m}}$ the reconstructed object appearances (RGB values for each pixel) and shapes (foreground probabiliy for each each pixel). The reconstruction loss $\mathcal{L}_R$ for these objects is then the weighted sum of the pixel-wise MSE for the appearance and the pixel-wise binary cross entropy for the mask:

$$\mathcal{L}_{R,\text{appear.}} = \sum_i \Big(\sum_{u,v} \mathbf{m}_1^{(i)}(u,v) \left\| \mathbf{c}^{(i)}(u,v) - \hat{\mathbf{c}}^{(i)}(u,v) \right\|_2^2 \Big/ \sum_{u,v} \mathbf{m}_1^{(i)}(u,v) \Big),$$

$$\mathcal{L}_{R,\text{mask}} = \sum_i \Big(\sum_{u,v} [\mathbf{m}_0^{(i)} + \mathbf{m}_1^{(i)}](u,v) \cdot \text{BCE}\big[\mathbf{m}_1^{(i)}(u,v), \hat{\mathbf{m}}^{(i)}(u,v)\big] \Big/ \sum_{u,v} [\mathbf{m}_0^{(i)} + \mathbf{m}_1^{(i)}](u,v) \Big).$$

As the task for the object model is to only represent the central object in each crop, we restrict the appearance loss to the candidate mask of the object ($\mathbf{m}_1$) and the mask loss to the union of the candidates masks of the object and the background ($\mathbf{m}_0 + \mathbf{m}_1$). Importantly, the reconstruction loss is not evaluated for pixels belonging to other objects according to the candidate masks. Therefore, the object model is not penalized for completing object parts that are occluded by another object.

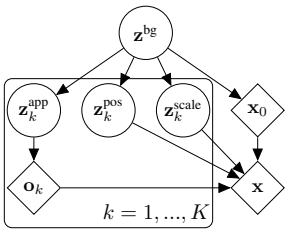

Figure 3: Causal graph for our scene model; circles and diamonds denote random and deterministic quantities.

Table 1: Segmentation performance of the unsupervised motion segmentation (Keuper et al., 2015) for different optical flow estimators.

Optical flow		**IoU**		**Recall**	
Estimator	Training data	Background	Objects	@0.0	@0.5
ARFlow	KITTI	0.874	0.246	0.828	0.199
ARFlow	Sintel	0.890	0.243	0.809	0.213
ARFlow	Fishbowl	0.873	0.248	0.842	0.204
RAFT	FlyingThings	0.930	0.318	0.663	0.351
FlowNet 2	FlyingThings	**0.934**	**0.365**	**0.674**	**0.416**

Learning object completion via artificial occlusions. To encourage the model to correctly complete partial objects, we use artificial occlusions as an augmentation during training. Similar to a denoising autoencoder (Vincent et al., 2008), we compute the reconstruction loss using the unaugmented object crop. We consider two types of artificial occlusions: first, we use a *cutout* augmentation (DeVries and Taylor, 2017) placing a variable number of grey rectangles on the input image. As an alternative, we use the candidate segmentation to place another, randomly shifted object from the same input batch onto each crop.

Model. We use a β-VAE with 128 latent dimensions. The encoder is a ten layer CNN, the appearance decoder is a corresponding CNN using transposed convolutions (Dumoulin and Visin, 2018) and one additional convolutional decoding layer. We use a second decoder with the same architecture but only a single output channel do decode the object masks. During each epoch, we use crops from two random frames from every object. We train our model for 60 epochs using Adam (Kingma and Ba, 2015) with a learning rate of 10^{-4}, which we decrease by a factor of 10 after 40 epochs. We chose the optimal hyperparameters for this architecture using grid searches. More details regarding the model architecture and the hyperparameters are provided in the supplement.

STAGE 2B: BACKGROUND MODEL—LEARNING TO GENERATE UNOCCLUDED BACKGROUNDS

Task. We use an ensemble of background extraction techniques outlined above to estimate background scenes for each frame. We train a β-VAE on these backgrounds using the appearance loss $\mathcal{L}_{R,\text{appear.}}$ with the inferred background mask, without any additional cutout or object augmentation.

Architecture. The β-VAE has the same architecture as the object model, but only uses a single decoder for the background appearance. We do not focus on a detailed reconstruction of background samples and limit the resolution to 96×64px. When sampling scenes, the outputs are upsampled to the original resolution of 480×320px using bilinear interpolation.

STAGE 3: SCENE MODEL—LEARNING TO GENERATE COHERENT SCENES

In the final stage, we combine the object and background model into a scene model that allows sampling novel scenes. As the scene model can reuse the decoders from the previous stages, its main task is to model the parameters defining the scene composition such as object counts, locations and dependencies between the background and the object latents. Compared to an end-to-end approach, the complexity of the learning problem is greatly reduced in this setting. It is straightforward to

generalize the scene model beyond the training distribution: E.g., it is easy to sample more objects than observed in the input scenes.

We use a scene model following the causal graph depicted in Fig. 3: First, we sample a background latent $\mathbf{z}^{\text{bg}}$ which describes global properties of the scene such as its composition and illumination; $\mathbf{z}^{\text{bg}}$ is then decoded by the background model into a background image $\mathbf{x}_0$. Conditioned on the background latent, we sequentially sample K tuples $(\mathbf{z}_k^{\text{app}}, \mathbf{z}_k^{\text{pos}}, \mathbf{z}_k^{\text{scale}})$ of latents encoding appearance, position, and scale of object k, respectively; the number of objects K is sampled conditional on $\mathbf{z}^{\text{bg}}$ as well. Each appearance latent $\mathbf{z}_k^{\text{app}}$ is decoded by the object model into a masked object $\mathbf{o}_k = (\mathbf{m}_i, \mathbf{x}_i)$, which is subsequently re-scaled by $\mathbf{z}_k^{\text{scale}}$ and placed in the scene at position $\mathbf{z}_k^{\text{pos}}$ according to a dead-leaves model (i.e., occluding previously visible pixels at the same location).

Due to the formulation of the model, we are flexible in specifying the conditional and prior distributions needed to generate samples. A particular simple special case is to sample all latents (indicated as circles in Fig. 3) independently. This can be done by informed prior distributions, or by leveraging the training dataset. In the former case, we sample $\mathbf{z}^{\text{bg}}$ and all $\mathbf{z}_k^{\text{app}}$ from the standard normal prior of the β-VAE, but reject objects for which the binary entropy of the mask (averaged across all pixels) exceeds a threshold (for figures in the main paper, 100 bits). We found that empirically, this entropy threshold can be used to trade diversity of samples for higher-quality samples (cf. supplement). For the coordinates, a uniform prior within the image yields reasonable samples, and scales can be sampled from a uniform distribution between 64×32px and 192×96px and at fixed $2:1$ ratio. Alternatively, all distributions can be fit based on values obtained from the motion segmentation (object and background latents, distribution of sizes, distribution of coordinates). We provide a detailed analysis in the supplement.

5. Experiments

Motion Segmentation. We quantify the quality of the motion segmentation approach to understand which and how many errors are propagated to later stages. Motion segmentation was scored as follows: For every frame, we matched the predicted and ground truth masks using the Hungarian algorithm (Kuhn, 1955) with the pairwise IoU (Intersection over Union) as matching cost. The frame-wise IoU scores were then aggregated as in the DAVIS-2017 benchmark (Pont-Tuset et al., 2017) by first averaging IoUs over frames for each ground truth object, and then averaging over objects. In Tab. 1 we report the segmentation performance separately for the background and the foreground objects.

The best performance is reached when using optical flow predicted by the FlowNet 2 model, thus we're using this model for all subsequent experiments. Our setting seems to be especially difficult for the ARFlow model. Different from the other networks, this model does not easily transfer from unrelated training settings and doesn't improve much when training on the Fishbowl data directly using the default hyperparameters. It might however be possible to improve the model by adapting the training setup to our dataset more closely.

Overall, the evaluation reveals two types of errors: (1) even the best model variant only detected 67.4% of all objects, i.e., one third of the objects are missed and get included in the background masks. (2) The recall decreases when increasing the IoU threshold to 0.5, i.e., the predicted masks are often not precise. While the motion segmentation therefore induces a sort of label noise, it has to be considered that not all errors are problematic for the motion segmentation. In particular, this holds for all objects that have not been detected at all and unsystematic errors.

Table 2: Comparison of the reconstructions from the object model with the ground truth unoccluded objects. The reconstructed masks are evaluated using the intersection over union (IoU), the appearance is evaluated using the mean average error (MAE) of the RGB values (0–255) on the intersection of the ground truth and predicted masks. Additionally, we report those metrics for hard samples in which at most half of the object is visible (IoU@0.5 and MAE@0.5). We consider model variants that use different artificial occlusions as augmentations during training. Furthermore, we compare the performance of each model variant when trained on the motion segmentation and the ground truth masks, respectively. All results are averaged over independent runs with 3 different seeds and reported with the standard error of the mean. As baseline we report the IoU of the ground truth occluded masks with the ground truth unoccluded masks.

Training data	**Augmentation**	**IoU** ↑	**MAE** ↓	**IoU@0.5** ↑	**MAE@0.5** ↓
Motion Segmentation	None	0.820±0.002	13.0±0.047	0.669±0.002	24.3±0.037
	Cutout	0.822±0.001	**13.0±0.032**	0.677±0.002	24.0±0.017
	Other object	**0.827±0.001**	14.8±0.095	**0.705±0.001**	**23.4±0.049**
Ground Truth Segmentation	None	0.885±0.001	12.4±0.144	0.738±0.002	17.5±0.210
	Cutout	**0.887±0.001**	**12.3±0.035**	0.743±0.003	**17.3±0.087**
	Other object	0.883±0.001	14.1±0.250	**0.782±0.002**	17.7±0.197
Baseline		0.915		0.271	

Object model. To evaluate the capability of the object model to reconstruct complete objects from occluded inputs, we extract object crops from the validation set using the bounding boxes of the ground truth unoccluded masks. For those crops, we compare the masks and appearances reconstructed by the model to the unoccluded ground truth (Fig. 2) using IoU and mean average error (MAE), respectively. As we are only interested in the reconstruction error within the object masks, we evaluate the MAE only on the intersection of the predicted and ground truth mask. We consider the ground truth occluded segmentation masks as our baseline: they correspond to a model that perfectly segments all visible object parts, but does not complete partial objects. We train variants of the object model using all augmentation strategies described above. For comparison, we also train each model variant using the ground truth occluded segmentation masks so that errors propagated from the motion segmentation can be differentiated from errors inherent to the object model.

As the results in Tab. 2 show, using another object from the input batch as artificial occlusions results in the best IoU of 0.827. When considering only those objects that are occluded at least by 50%, the IoU drops to 0.705. This performance is substantially greater than the baseline (0.27), which indicates that the model learns to complete partial objects even for those hard cases. In terms of the appearance error, using the cutout augmentation performs best. Interestingly though, for both the mask and appearance error, training without any augmentation at all reduces the performance only slightly.

When training on the ground truth segmentation instead of the motion segmentation, the error in terms of the IoU reduces by roughly one third. Performance gains can therefore be expected from improving the generative object model and, to a lesser extent, the motion segmentation. Overall, the object model seems to be quite robust to the label noise induced by the motion segmentation.

We visualize reconstructions and samples from the object model in Fig. 4. In agreement with the quantitative results before, there is a visible quality difference between the model trained on

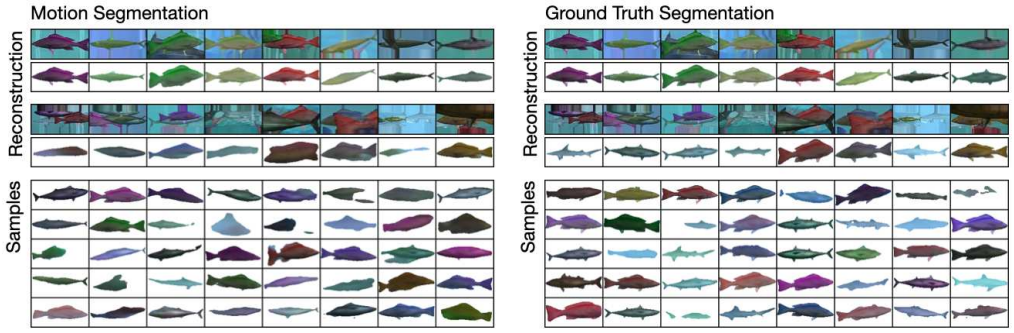

Figure 4: Qualitative results from the object model using another object as augmentation. *Left:* Object model trained on the masks predicted by motion segmentation. *Right:* Object model trained using ground truth segmentations. *Top:* Reconstructions of validation set elements. *Middle:* Reconstructions of validation set elements that are occluded by 0.5 ± 0.05. *Bottom:* objects sampled from the respective model. The shown reconstructions and samples are not cherrypicked and do not use entropy filtering as in the scene model.

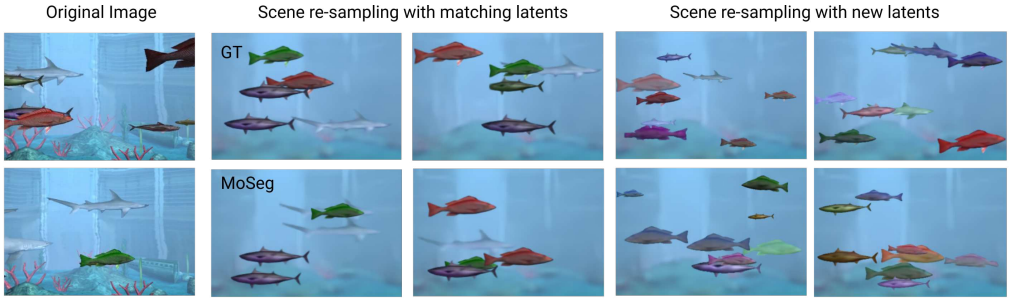

Figure 5: Samples from the training dataset in comparison to the scene model using the background and object models trained on the motion segmentation and ground truth segmentation, respectively.

motion segmentation and ground truth segmentation, respectively. In some cases, parts of the fish such as the tail fin is missing and in other cases, the mask covers areas not properly covered by the appearance. In particular this holds for the hard cases with substantial occlusions. Moreover, the samples from the model trained on motion segmentation masks seems to contain fewer sharks (a rare class). The majority of the reconstructions and samples however looks convincing, with the model capturing the relevant properties of the object masks and appearances.

Background model. We train two variants of the background model using the foreground/background segmentation and the ground truth segmentation masks, respectively. We show samples from the former variants as part of the scene model in Fig. 5, and provide additional samples in the supplement. We deliberately chose to not exhaustively tune the background model and defer improvements (which are conceivable given the recent advances in high resolution image modeling with VAEs) to future work. Within the scope of this work, we limit ourselves to the constrained resolution resulting in blurry samples. They still reflect the main variations in the dataset, like the overall illumination and background color.

Scene model. In Fig. 5 we visualize samples from the scene model following the scene statistics from the training dataset. As before, we consider two variants of the scene model using the

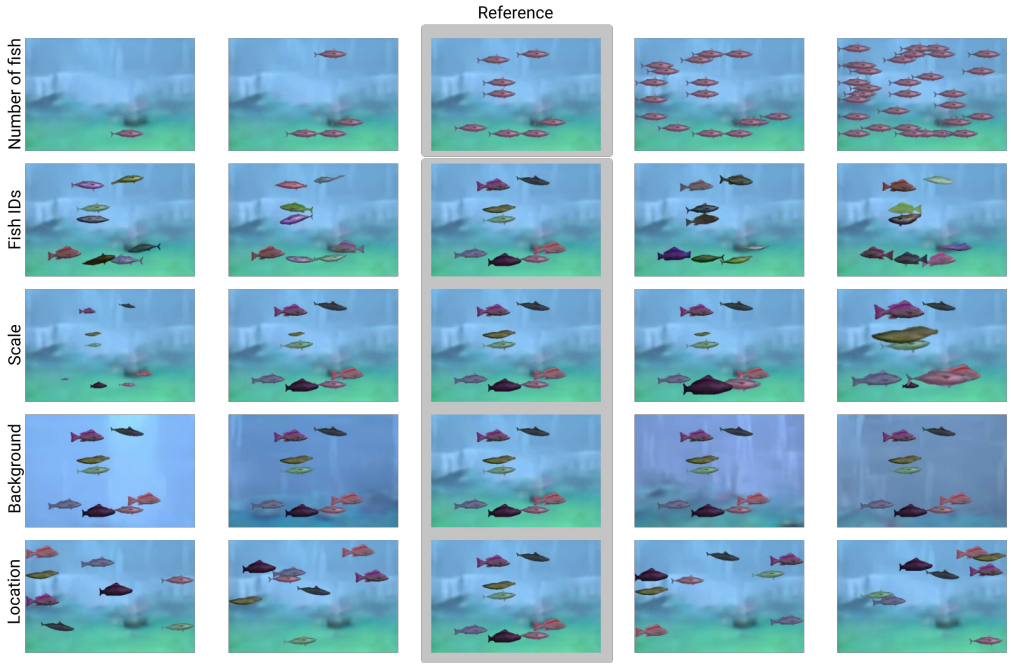

Figure 6: Interventions on the scene properties of a sample from the unsupervised scene model. Top to bottom: (i) Varying the number of objects, (ii) varying the object identities, (iii) varying the scales of individual objects, (iv) changing backgrounds while keeping objects constant, (v) re-sampling locations of all objects.

background and object models trained on the motion segmentation and ground truth segmentation, respectively. The samples from both models nicely resemble the scenes from the training dataset. However, in both cases the errors from the object and the background model, as discussed before, are also observable in the generated scenes.

A particular strength of our modular approach is that it allows to separately intervene on causally independent mechanisms (Pearl, 2009; Schölkopf et al., 2021) of the scene generation. In Fig. 6 we show such interventions on a sample from the unsupervised scene model. The number of objects, their positions and sizes are all represented explicitly by the scene model so that we can manipulate those scene properties separately. Moreover, the background and the objects are represented separately by their respective latent codes so that we can exchange and manipulate each of those scene components individually.

Modularity. A downside of end-to-end models are non-trivial dependencies between individual components, making it harder to train and debug such models. In contrast, our modular approach allows to exchange individual components in the object learning pipeline. In Figure 7, we show a proof-of-concept for using a GAN (Mescheder et al., 2018) as part of a hybrid object model (image modeling: GAN, mask modeling: VAE) within our framework.

5.1. Comparison to prior work

The previous experiments have shown that our modular object learning approach scales well to the visually more complex FISHBOWL dataset. In the following, we compare our method with several

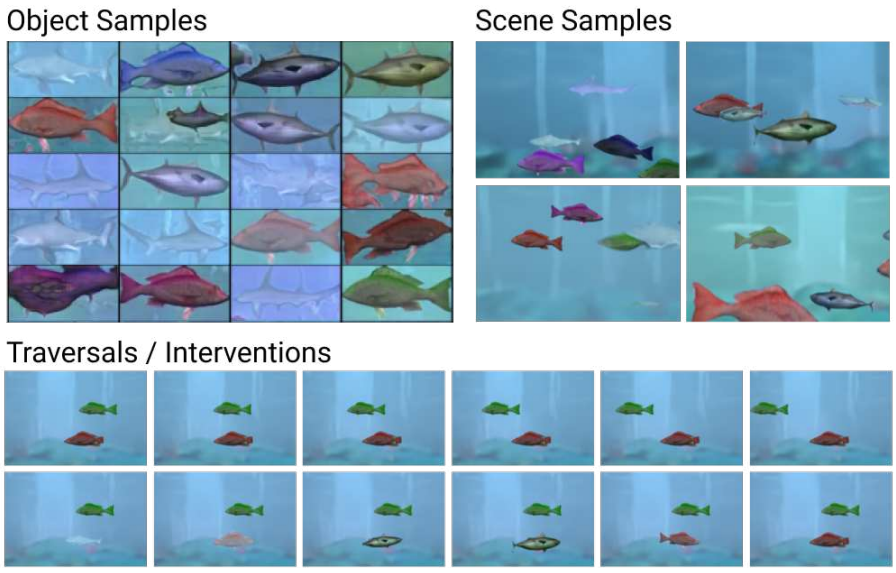

Figure 7: The modularity of our multi-stage approach allows to easily exchange individual components. Here, we replace the image-modeling part of the object model with a GAN instead of the previously used VAE. As in our main model, this allows meaningful traversals/interventions since objects and their position-s/sizes are presented as individual entities.

established approaches for object learning and scene generation. In each case, we summarize the key findings and provide a more detailed treatment in the appendix.

A key feature of our model is the disentanglement of an object's appearance and its location in the scene. Several prior works approached learning such disentangled object models end-to-end. We compare to SPACE (Lin et al., 2020b) as a representative method. SPACE includes a mechanism for modelling occlusions, allowing it to model complete objects as in our approach. We find that SPACE fails to learn the correct object notion on the FISHBOWL dataset. This is in line with previous results showing that previous approaches do not scale well to textured data (Karazija et al., 2021). When trained with our approach, the same object model as used by SPACE learns the objects much better.

Another line of works doesn't entangle object appearance and position, but rather learns a layered scene representation with each layer covering the entire image. We compared to GENESIS-v2 (Engelcke et al., 2021) as a recent, probabilistic variant of this approach. When trained with the GECO objective as proposed by the authors, GENESIS-v2 learns to detect objects but fails at generating novel scenes. When using the original VAE objective, sampling from the model works but objects are not recovered. Again, this is in line with previous work showing that object-centric methods struggle on textured data (Karazija et al., 2021).

GANs (Goodfellow et al., 2014) are a powerful method for generating images. We trained the variant by Mescheder et al. (2018) on the Fishbowl dataset. Overall, the GAN is able to generate images of convincing quality resembling the training images well. A particular strength of the GAN in comparison to our method, is it's ability to generate backgrounds with many details—which we explicitly left for future work. Compared to our method, the GAN however lacks a principled way to manipulate sampled scenes and to generalize beyond the training distribution.

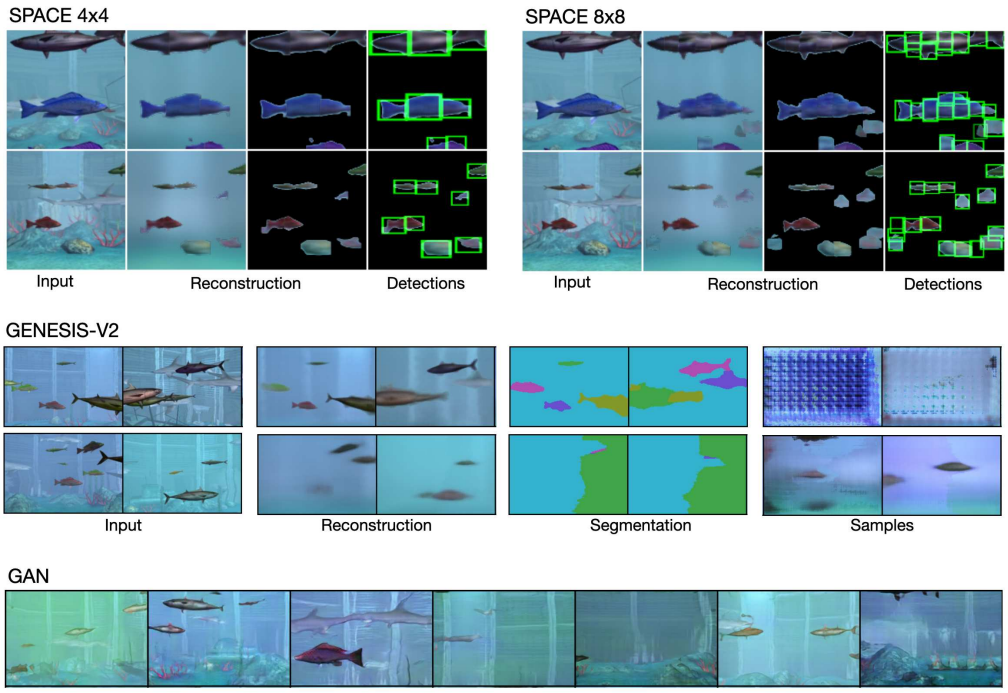

Figure 8: Qualitative comparison with other models trained on the FISHBOWL dataset. Top: SPACE (Lin et al., 2020b), using 4x4 or 8x8 grid cells. Middle: GENESIS-v2 (Engelcke et al., 2021) trained both using the GECO (top) and the original VAE objective (bottom). Bottom: GAN following Mescheder et al. (2018).

To evaluate how well our method transfers to other settings, we trained our model on the RealTraffic dataset (Ehrhardt et al., 2020). In Fig. 26, we compare samples from our model to other models for which results on the RealTraffic dataset have been reported by Ehrhardt et al. (2020). Overall, our model transfers well to this real world setting, given that our model is used largely unchanged. In comparison to the GAN-based RELATE and BlockGAN, the samples from our model look slightly more blurry. However the results from our method clearly improve over the VAE-based GENESIS model.

6. Discussion

The fields of machine learning and causal modeling have recently begun to cross-pollinate. Causality offers a principled framework to think about classes of distributions related to each other by domain shifts and interventions, and independent/modular components (e.g., objects) that are invariant across such changes. Vice versa, machine learning has something to offer to causality when it comes to *learning* causal variables and representations—traditionally, research in causal discovery and inference built on the assumption that those variables be given a priori. Much like the shift from classic AI towards machine learning included a shift from processing human-defined symbols towards automatically learning features and *statistical* representations, causal representation learning aims to learn *interventional* representations from raw data (Schölkopf et al., 2021). It is in this sense that we like to think of our scene model as a causal generative model. Its causal structure is simple—its complexity lies on the side of representation learning.

We propose a multi-stage approach for learning the model from unlabelled videos. Inspired by the Gestalt *Principle of Common Fate*, we use candidate segmentations obtained from unsupervised optic flow to weakly supervise generative object and background models, and combine them into a scene model. The scene model is causal/mechanistic in the sense that it (1) learns complete objects (although occlusions are abundant in the training data)[2] and (2) properly handles partial depth orderings in the scene, the latter through the "dead leaves" approach—as opposed to the computationally simpler additive mixture approach which is common place in end-to-end scene models, but contradicts the generative structure of scenes. Moreover, while many existing models only decompose and represent scenes, (3) our model can generate plausible novel and out-of-distribution scenes (e.g., higher fish density, different sizes).

While learning object models as part of an end-to-end trainable object centric scene model might have certain advantages, it has not been shown to scale to more realistic data yet. We instead chose a multi-stage approach that allows developing and analyzing the modeling stages individually. This approach allowed us to conclude that exploiting motion information for object discovery, as done in the motion segmentation stage, is a valuable cue that greatly simplifies object learning. Also, since the scene model uses the component models as modelling "atoms", this naturally permits intervening on the parameters defining the scene composition without side-effects, as stipulated by the principle of Independent Causal Mechanisms (Parascandolo et al., 2018; Peters et al., 2017).

We evaluated the modelling stages on the novel FISHBOWL dataset, consisting of short videos recorded from a synthetic 3D aquarium. Notably, we found that the generative object model proves to be fairly robust to errors in the motion segmentation. We expect further performance gains to be achievable by improving the visual fidelity of the generative object and background models (Razavi et al., 2019; Vahdat and Kautz, 2020); however, this was not the motivation for the present work.

We see several ways to extend our model in the future, beyond improving the implementations of the individual stages by taking advantage of progress in computer vision. One limitation of using motion segmentation is that we can only decompose scenes of moving objects. This could be addressed by integrating object models trained using our approach into an end-to-end scene model (similar to Bao et al., 2022). When using a learning-based motion segmentation approach, there could be a feedback circle with both models improving each other as training progresses and this way closing the observed gap to using ground truth segmentations. Also, our object model does currently not capture object motions. Given that the motion segmentation yields temporally consistent segmentation masks, this could, in principle, well be included in our approach. Finally, one could imagine also using interaction data where an agent intervenes in a scene and infers objects from how it responds to those interventions.

Acknowledgements

Part of this work was done while MT, StS, and JvK were interning at Amazon. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 4, project number 276693517, and Germany's Excellence Strategy – EXC 2064/1 – 390727645. It was also supported by the German Federal Ministry of Education and Research (BMBF), FKZ 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems for supporting MT and StS.

2. Inpainting takes place at the level of masks/shapes, rather than only for pixels as in most prior work.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer Normalization. *arXiv:1607.06450*, July 2016.

Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering Objects That Can Move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11789–11798, June 2022.

David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a GAN cannot generate. In *Proceedings ICCV*, pages 4502–4511, 2019.

S. Beucher and F. Meyer. The Morphological Approach to Segmentation: The Watershed Transformation. In E. R. Dougherty, editor, *Mathematical Morphology in Image Processing*, pages 433–481. Marcel Dekker, New York, 1993.

Thomas G. Bever and David Poeppel. Analysis by synthesis: A (re-)emerging program of research for language and vision. *Biolinguistics*, 4(2):174–200, 2010.

Thomas Brox and Jitendra Malik. Object Segmentation by Long Term Analysis of Point Trajectories. In *ECCV 2010*, pages 282–295, Berlin, Heidelberg, September 2010. Springer.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised Scene Decomposition and Representation. *arXiv:1901.11390*, 2019.

Chang Chen, Fei Deng, and Sungjin Ahn. Learning to infer 3d object models from images. *arXiv preprint arXiv:2006.06130*, 2020.

Honglin Chen, Rahul Venkatesh, Yoni Friedman, Jiajun Wu, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Daniel M. Bear. Unsupervised Segmentation in Real-World Images via Spelke Object Inference, May 2022.

Eric Crawford and Joelle Pineau. Spatial invariant unsupervised object detection with convolutional neural networks. In *Thirty-Third AAAI*, 2019.

Eric Crawford and Joelle Pineau. Exploiting Spatial Invariance for Scalable Unsupervised Object Tracking. In *Proceedings AAAI*, volume 34(04), pages 3684–3692, April 2020.

Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *Proc. IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

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

Andrea Dittadi and Ole Winther. LAVAE: Disentangling Location and Appearance. *arXiv:1909.11813*, September 2019.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet. In *Proceedings ICCV*, pages 2758–2766, December 2015.

Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv:1603.07285*, January 2018.

Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. RELATE: Physically Plausible Multi-Object Scene Synthesis. In *NeurIPS*, volume 33, 2020.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations. In *ICLR*, 2020.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. GENESIS-V2: Inferring unordered object representations without iterative refinement. *arXiv:2104.09958*, 2021.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E Hinton. Attend, Infer, Repeat: Fast Scene Understanding with Generative Models. In *NeurIPS 29*, pages 3225–3233, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS 27*, pages 2672–2680. 2014.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *NeurIPS*, pages 6694–6704, 2017.

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 *ICML*, volume 97, pages 2424–2433. PMLR, 2019.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the Binding Problem in Artificial Neural Networks. *arXiv:2012.05208*, 2020.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3761, June 2022.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *ICML*, pages 1462–1471, 2015.

Nicolas Heess, Nicolas Le Roux, and John Winn. Weakly supervised learning of foreground-background segmentation using masked RBMs. In *ICANN*, pages 9–16. Springer, 2011.

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 *ICLR*, 2017.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. In *CVPR*, pages 2462–2470, July 2017.

Varun Jampani, Sebastian Nowozin, Matthew Loper, and Peter V. Gehler. The informed sampler: A discriminative approach to Bayesian inference in generative computer vision models. *Computer Vision and Image Understanding*, 136:32 – 44, 2015.

Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative World Models with Scalable Object Representations. *arXiv:1910.02384*, October 2019.

Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A Texture-Rich Benchmark for Unsupervised Multi-Object Segmentation. *arXiv:2111.10265 [cs]*, November 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*, February 2018.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*, pages 8110–8119, 2020.

Margret Keuper, Bjoern Andres, and Thomas Brox. Motion Trajectory Segmentation via Minimum Cost Multicuts. In *ICCV*, pages 3271–3279, December 2015.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, May 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *ICLR 2022*, April 2022.

Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential Attend, Infer, Repeat: Generative Modelling of Moving Objects. In *NeurIPS*, volume 31, 2018.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: A probabilistic programming language for scene perception. In *CVPR*, pages 4390–4399, 2015.

Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23(3):593–650, 2011.

Zhixuan Lin, Yi-Fu Wu, Skand Peri, Bofeng Fu, Jindong Jiang, and Sungjin Ahn. Improving Generative Imagination in Object-Centric World Models. In *ICML*, pages 6140–6149. PMLR, 2020a.

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition. In *ICLR*, 2020b.

Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by Analogy: Reliable Supervision From Transformations for Unsupervised Optical Flow Estimation. In *CVPR*, pages 6489–6498, 2020.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *NeurIPS*, 2020.

Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Cross Pixel Optical Flow Similarity for Self-Supervised Learning. *arXiv:1807.05636*, 2018a.

Aravindh Mahendran, James Thewlis, and Andrea Vedaldi. Self-Supervised Segmentation by Grouping Optical-Flow. In *ECCV Workshops*, 2018b.

Antoine Manzanera and Julien C Richefeu. A new motion detection algorithm based on σ–δ background estimation. *Pattern Recognition Letters*, 28(3):320–328, 2007.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *CVPR*, pages 4040–4048, 2016.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs do actually Converge? In *ICML*, pages 3481–3490. PMLR, July 2018.

Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. EM-driven unsupervised learning for efficient motion segmentation, January 2022.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

Li Nanbo, Cian Eastwood, and Robert Fisher. Learning object-centric representations of multi-object scenes from multiple views. *Advances in Neural Information Processing Systems*, 33, 2020.

Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3d object-aware scene representations from unlabelled images. *arXiv:2002.08988*, 2020.

Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021.

Peter Ochs and Thomas Brox. Higher order motion models and spectral clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 614–621, June 2012.

Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(6):1187–1200, 2014. ISSN 0162-8828, 2160-9292.

G. Parascandolo, N. Kilbertus, M. Rojas-Carulla, and B. Schölkopf. Learning independent causal mechanisms. In *Proc. 35th ICML*, pages 4033–4041, 2018.

Deepak Pathak, Ross Girshick, Piotr Dollar, Trevor Darrell, and Bharath Hariharan. Learning Features by Watching Objects Move. In *CVPR*, pages 2701–2710, 2017.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, USA, second edition, November 2009. ISBN 978-0-511-80316-1.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

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:1704.00675*, 2017.

Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *CVPR*, pages 12240–12249, 2019.

Ali Razavi, Aäron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *NeurIPS*, volume 32, pages 14866–14876, 2019.

Danilo Jimenez Rezende and Fabio Viola. Taming VAEs. *arXiv:1810.00597*, October 2018.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286, 2014.

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.

William Silversmith. cc3d: Connected components on multilabel 3D images. https://github.com/seung-lab/connected-components-3d, GPL-3.0 License, January 2021.

Andrews Sobral. BGSLibrary: An OpenCV C++ background subtraction library. In *IX Workshop de Visão Computacional (WVC'2013)*, Rio de Janeiro, Brazil, 2013. https://github.com/andrewssobral/bgslibrary.

Elizabeth S. Spelke. Principles of object perception. *Cognitive Science*, 14(1):29–56, 1990. ISSN 0364-0213.

Pierre-Luc St-Charles and Guillaume-Alexandre Bilodeau. Improving background subtraction using local binary similarity patterns. In *IEEE Winter Conf. on Applic. of Computer Vision*, pages 509–515, 2014.

Pierre-Luc St-Charles, Guillaume-Alexandre Bilodeau, and Robert Bergevin. Universal background subtraction using word consensus models. *IEEE Transactions on Image Processing*, 25(10): 4768–4781, 2016.

Karl Stelzner, Robert Peharz, and Kristian Kersting. Faster Attend-Infer-Repeat with Tractable Probabilistic Models. In *ICML*, volume 97, pages 5966–5975. PMLR, 2019.

A.K Subramanian. Pytorch-vae. https://github.com/AntixK/PyTorch-VAE, 2020.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *The European Conference on Computer Vision (ECCV)*, August 2020.

Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning Video Object Segmentation With Visual Memory. In *ICCV*, pages 4481–4490, October 2017a.

Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning Motion Patterns in Videos. In *CVPR*, pages 3386–3394, 2017b.

Mehmet Ozgur Turkoglu, William Thong, Luuk Spreeuwers, and Berkay Kicanaoglu. A layer-based sequential framework for scene generation with gans. In *Proceedings AAAI*, volume 33, pages 8901–8908, 2019.

Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *ICLR*, 2018.

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 RL. In *CoRL*, 2019.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008.

Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. In *ICLR 2020 Workshops*, April 2020.

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.

Max Wertheimer. *On Perceived Motion and Figural Organization*. MIT Press, Cambridge, MA, 2012. New English translation of two articles from 1912/1923.

Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 699–707, 2017.

Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object Discovery in Videos as Foreground Motion Clustering. In *CVPR*, pages 9994–10003, 2019.

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised Video Object Segmentation by Motion Grouping. *arXiv:2104.07658*, April 2021.

Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-GAN: Layered recursive generative adversarial networks for image generation. In *ICLR*, 2017.

Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised Moving Object Detection via Contextual Information Separation. In *CVPR*, pages 879–888, June 2019.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *ICLR*, 2019.

Jinyang Yuan, Bin Li, and Xiangyang Xue. Generative modeling of infinite occluded objects for compositional scene representation. In *ICML*, pages 7222–7231, 2019.

Appendix A. Additional details about the method

A.1. Motion segmentation

We use the original implementation of the motion segmentation by Keuper et al. (2015), but replace the postprocessing required to obtain a dense segmentation with a simpler but faster non-parametric watershed algorithm (Beucher and Meyer, 1993) followed by computing spatiotemporal connected components (Silversmith, 2021).

To obtain background masks for training the background model, we aim for a low rate of foreground pixels being mistaken for background pixels, while background pixels mistaken for foreground are of less concern. Hence, we use an ensemble of different background-foreground segmentation models from the bgslibrary (Sobral, 2013). Based on early experiments, we used the PAWCS (St-Charles et al., 2016), LOBSTER (St-Charles and Bilodeau, 2014), $\Sigma - \Delta$ estimation (Manzanera and Richefeu, 2007) and static frame differences and label every pixel detected as foreground by either of the methods as a foreground pixel. We found that this rather simple model can faithfully remove the foreground objects in most of the cases.

A.2. Optical flow estimation

The quality of the motion segmentation critically depends on the quality on the optical flow estimation, so we explore different models for that step. ARFlow (Liu et al., 2020) is a state of the art self-supervised optical flow method that combines a common warping based objective with self-supervision using various augmentations. We use the published pretrained models as well as a variant trained on the Fishbowl dataset.

To train ARFlow on the Fishbowl dataset, e build on the official implementation provided in https://github.com/lliuz/ARFlow. We make the following adaptations to the original configuration used for training on Sintel:

- The internal resolution of the model is set to 320x448 pixels.
- For training, we use the first 200 videos of the Fishbowl dataset. This amounts to 25.4K frame pairs which is substantially more than the Sintel dataset (pretraining: 14,570, main training: 1041), which is the largest dataset on which ARFlow was originally trained. Initial experiments using the first 1000 videos did not lead to an improvement in the training objective compared to only using 200 videos.
- We train the model using a batch size of 24 and use 300 random batches per epoch. We perform both the pretraining and main training stage, but using the same data for both stages. The pretraining stage is shortened to 100 epochs, after which the training loss did not improve further.

We selected above parameters by pilot experiments using the final training loss as criterion. All other hyperparameters, in particular regarding the losses, the augmentations and the model architecture, are used unchanged.

We remark that the hyperparameters are chosen differently for the two datasets used in the original paper, so we conjecture that the performance on this model most likely improves when closely adapting the training scheme to our setting.

Similar augmentations as used by ARFlow can alternatively be used to synthesize training data for supervised methods, as done for generating the FlyingChairs and FlyingThings datasets (Dosovitskiy et al., 2015; Mayer et al., 2016). We experiment with FlowNet 2.0 (Ilg et al., 2017) and the more recent RAFT (Teed and Deng, 2020) trained on those two datasets.

A.3. Object model

We loosely build on the β-VAE implementation by Subramanian (2020). We reimplemented the training script and modified architecture details like the latent dimensionality due to differences in the image size.

We use a CNN with 10 layers as an encoder. Each layer consists of a 3×3 convolution, followed by layer normalization (Ba et al., 2016) and a leaky ReLU nonlinearity with a negative slope of 0.01. The decoders are built symmetrically by using the reversed list of layer parameters and transposed convolutions. The decoders both use an additional convolutional decoding layer, without normalization and nonlinearity. The detailed specification of the default hyperparameters used by the object model is given in the following table:

Table 3: Default hyperparameters used for the object model.

Parameter	Value
sample size	128×64
hidden layers: channels	32, 32, 64, 64, 128, 128, 256, 256, 512, 512
hidden layers: strides	2, 1, 2, 1, 2, 1, 2, 1, 2, 1
latent dimensions	128
prior loss weight (β)	0.0001
mask loss weight (γ)	0.1
learning rate, epoch 1-40	0.0001
learning rate, epoch 41-60	0.00001

We chose the parameters regarding the architecture based on early experiments by qualitatively evaluating the sample quality. The learning rate and mask loss weight γ where determined using grid searches with the IoU of the mask as selection criterion. However, we noticed a high degree of consistency between the rankings in terms of the mask IoU and appearance MAE. The best reconstruction quality was obtained when not regularizing the prior (i.e., $\beta = 0$). We chose the final value of $\beta = 0.0001$ as a compromise between reconsutruction and sampling capabilites based on visual inspection of the results.

For the mask and appearance losses defined in the main paper, we use an implementation based on the following PyTorch-like pseudo code:

```
def object_model_loss(image, mask_fg, mask_bg, gamma = 1., beta = 0.001):
  # image (b,c,h,w), mask_fg (b,h,w), mask_bfg (b,h,w)

  latents = encode(image)
  mask_pred, img_pred  = mask_decoder(latents), img_decoder(latents)

  L_img  = (mask_fg * mse(img_pred, image)).sum()
  L_mask = ((mask_fg + mask_bg) * bce(mask_pred, mask_fg)).sum()
  L_reg  = kl_divergence(latents)
  Z_img, Z_mask = mask_fg.sum(), (mask_fg + mask_bg).sum()
```

```
11
12    return L_img / Z_img + gamma * L_mask / Z_mask + beta * L_reg
```

A.4. Background model

Training details We use a similar β-VAE and training objective for the background model as for the object model. Different to the object model, we do not predict the mask and instead only reconstruct a non-occluded background sample. We sweep over learning rates in $\{1 \cdot 10^{-3}, 5 \cdot 10^{-4}, 1 \cdot 10^{-4}\}$ and β in $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-3}\}$ and select the model with $\beta = 10^{-3}$ and learning rate 10^{-4} which obtained lowest reconstruction and prior loss. A higher value for β caused the training to collapse (low prior loss, but high reconstruction loss). As noted in the main text, the background model performance (and resolution) can be very likely improved by using larger models, more hyperparameter tuning and better decoders suited to the high resolution. As noted in the main paper, we omit these optimizations within the scope of this work and rather focus on the object and scene models.

Implementation The background model loss is similar to the reconstruction loss of the foreground model. We omit reconstructing the mask. In the input image for the background model, all foreground pixels are replaced by the average background RGB value to avoid distorting the input color distribution, refer to the example images in Fig. 12. The loss can be implemented as follows:

```
def background_model_loss(image, mask_bg, beta = 0.001):
  # image (b,c,h,w), mask_fg (b,h,w), mask_bfg (b,h,w)

  latents  = encode(image)
  img_pred = img_decoder(latents)

  L_img = (mask_bg * mse(img_pred, image)).sum()
  L_reg = kl_divergence(latents)
  Z_img = mask_bg.sum()

  return L_img / Z_img + beta * L_reg
```

A.5. Scene model

Mask temperature By default, the mask is computed by passing the logits obtained from the object mask decoder through a sigmoid non-linearity for obtaining probabilities. During scene generation, we added sharpness to the samples by varying the temperature τ, yielding an object mask

$$\mathbf{m} = \frac{1}{1 + \exp(-\mathbf{x}/\tau)}, \tag{1}$$

where $\mathbf{x}$ is the logit output by the mask decoder and the exponential is applied element-wise. "Cooling" the model to values around $\tau = 0.1$ yields sharper masks and better perceived sample quality. Note that the entropy threshold introduced in the next section depends on the value of τ.

Entropy filtering When sampling objects during scene generation, we filter the samples according to the model "confidence". A simple form of rejection sampling is used by computing the mask for a given object latent, and then computing the entropy of that mask. Given a mask $\mathbf{m}(\tau)$, we

adapt the sampling process considering the average entropy across all pixels in the mask,

$$H_2(\mathbf{m}) = -\frac{1}{WH}\sum_{u=1}^{W}\sum_{v=1}^{H}\mathbf{m}(u,v)\log_2\mathbf{m}(u,v) + (1-\mathbf{m}(u,v))\log_2(1-\mathbf{m}(u,v)) \tag{2}$$

and reject samples where $H_2(\mathbf{m})$ exceeds a threshold. Reasonable values for a mask temperature of $\tau = 0.1$ are around 100 to 200 bits. It is possible to trade sample quality and sharpness for an increased variability in the samples by increasing the threshold.

Appendix B. Additional details about the dataset generation

A demo of the original aquarium WebGL demo used for our experiments is available at webglsamples.org/aquarium/aquarium.html. The original source code is released under a BSD 3-clause license (Google Inc.) and available at github.com/WebGLSamples/WebGLSamples.github.io.

We apply the following modifications to the original demo for our recording setup:

- We implemented a client-server recording setup: The client pulls recording parameters from the server and sends back the recorded frames with ground truth segmentation to the server. The server maintains the overall list of randomly sampled aquarium configurations and post-processes the recordings into their final form. We run this setup in parallel on multiple nodes, using the headless Chrome browser[3].
- All samples are recorded for 128 frames at 30Hz. We modified the aquarium to only advance to the next frame once the current frame is captured and sent to the server.
- For recording the masks, we modified the texture shaders to use a unique color for rendering every object. Furthermore, we disabled all rendering steps modifying those colors (anti-aliasing, alpha compositing).
- To make the recording fully deterministic, we disabled the bubbles in the aquarium.
- To increase the diversity of the fish in the aquarium, we applied random color shifts to all fish textures except sharks, for which those color shifts look very unnatural. We shifted each color by rotating each color value in the IQ plane of the YIQ color space. For each fish, we independently and uniformly sampled one of 8 discrete color shifts regularly spaced in $[0°, 360°]$.
- During post-processing, we used ffmpeg[4] to rescale and centrally crop 480×320px from every frame and encode all frames as a video. We cropped and rescaled the masks same way, but exported them as json using the run length encoding provided by the COCO API[5].
- For recording the unoccluded ground truth, we used above setup to record and post-process frames and masks of the validation and test samples, but only render the respective fish of interest.

The code used to generate the dataset including all parameters will be made publicly available.

3. https://github.com/puppeteer/puppeteer/
4. https://ffmpeg.org
5. https://github.com/cocodataset/cocoapi

Appendix C. Additional results

C.1. Objects extracted by the motion segmentation

The following figure shows examples of objects extracted from the original videos using motion segmentation and the ground truth occluded masks, respectively. The figure reveals typical failure modes, such as multiple fish that are segmented jointly, and parts of the background contained in the object mask. Most of the fish however, are segmented very accurately.

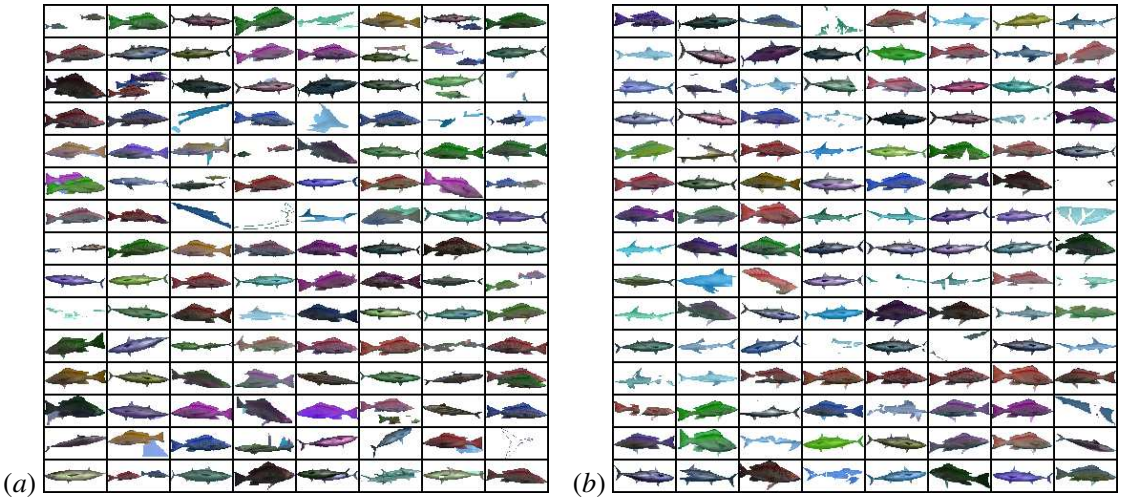

Figure 9: Objects extracted from the training videos that are used for training the object model. *Left:* Objects extracted using the motion segmentation. *Right:* Objects extracted using the ground truth occluded segmentation masks.

C.2. Object model: additional reconstructions

In Fig. 10 we show additional reconstruction from the object model trained on motion segmentation and ground truth occluded masks, respectively. Moderate occlusion levels up to 30% are handled well by both variants. At higher noise levels however, only the variant of the object model trained on the ground truth masks is able to correctly complete the partial objects.

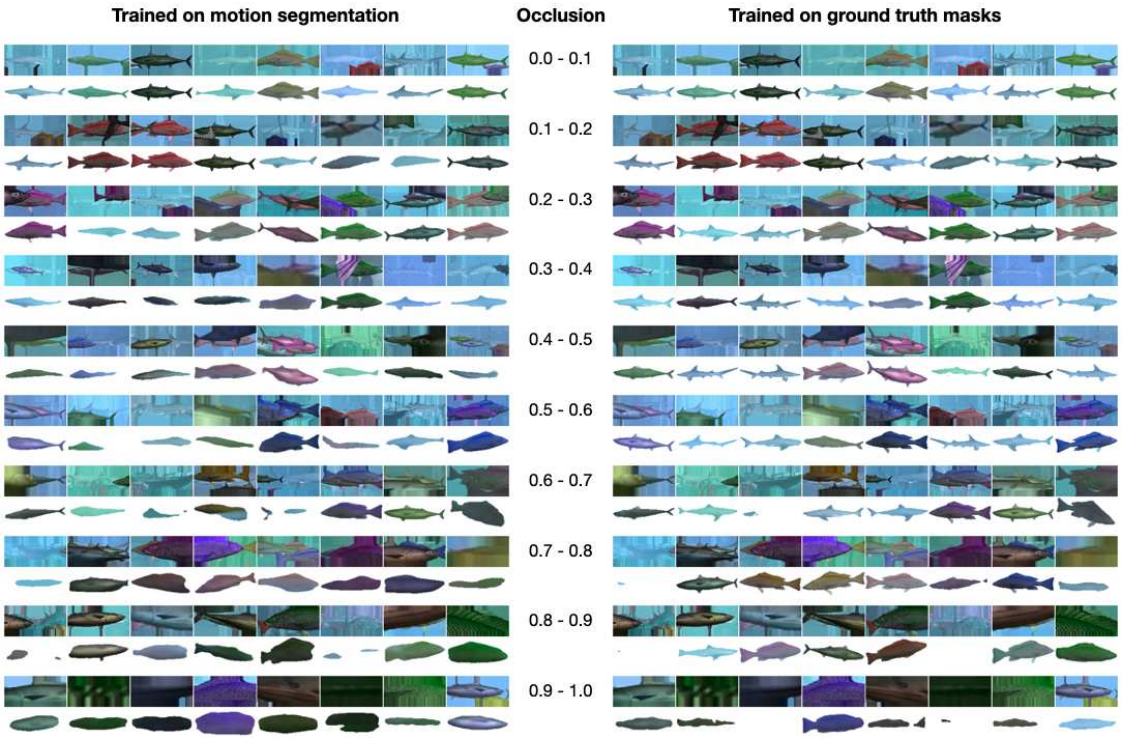

Figure 10: Reconstructions from the object model for input crops with different occlusion levels (0.0 = no occlusion, 1.0 = fully occluded). For each occlusion level, the input images and the respective reconstructions are shown. Both model variants are trained using another fish as artificial occlusion during training. *Left:* Object model trained using the motion segmentation masks. *Right:* Object model trained using the ground truth unoccluded masks.

C.3. Object model: additional samples

The following figure shows additional samples from the object models trained using another object as artificial occlusion (the same models as used for Fig. 4 in the main paper).

Figure 11: Samples from the object model using another input object as augmentation during training. These are the same models as used for Fig. 4 in the main paper. *Left:* Object model trained on the motion segmentation. *Right:* Object model trained on the ground truth occluded masks.

C.4. Background Model: Inputs

The background model is trained on images pre-processed by an ensembling of foreground-background algorithms from Sobral (2013). We use $\Sigma - \Delta$ (Manzanera and Richefeu, 2007), static frame dif-

ferences, LOBSTER (St-Charles and Bilodeau, 2014) and PAWKS (St-Charles et al., 2016) with standard hyperparameters set by Sobral (2013). The goal behind this choice is to detect as many objects as possible and remove the amount of erroneously included foreground pixels — it might be possible to even further improve this pre-processing step with different algorithms and hyperparameter tuning.

We give a qualitative impression of the background input samples in Fig. 12. Note that all foreground pixels were replaced by the average color value obtained by averaging over background pixels to avoid inputting unrealistic colors into the β-VAE.

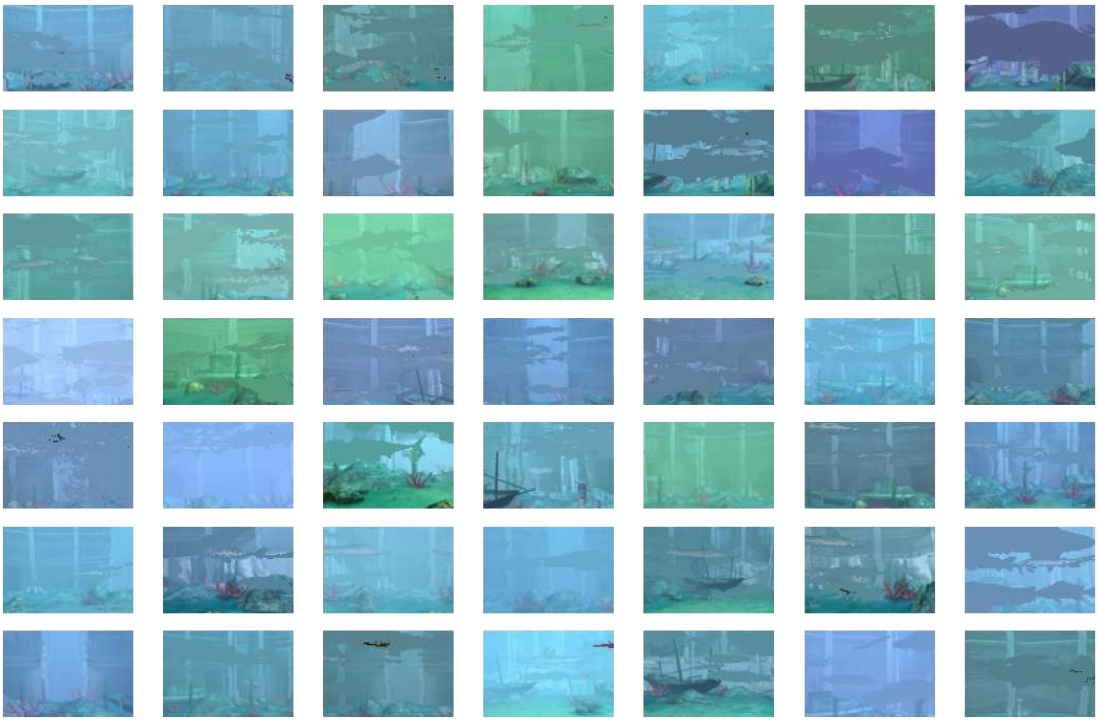

Figure 12: Input samples to the background β-VAE. Objects were removed by applying an ensemble of foreground-background segmentation algorithms. Images are resized to 96×64px to trade-off training speed for sample quality.

C.5. Additional samples from the scene model

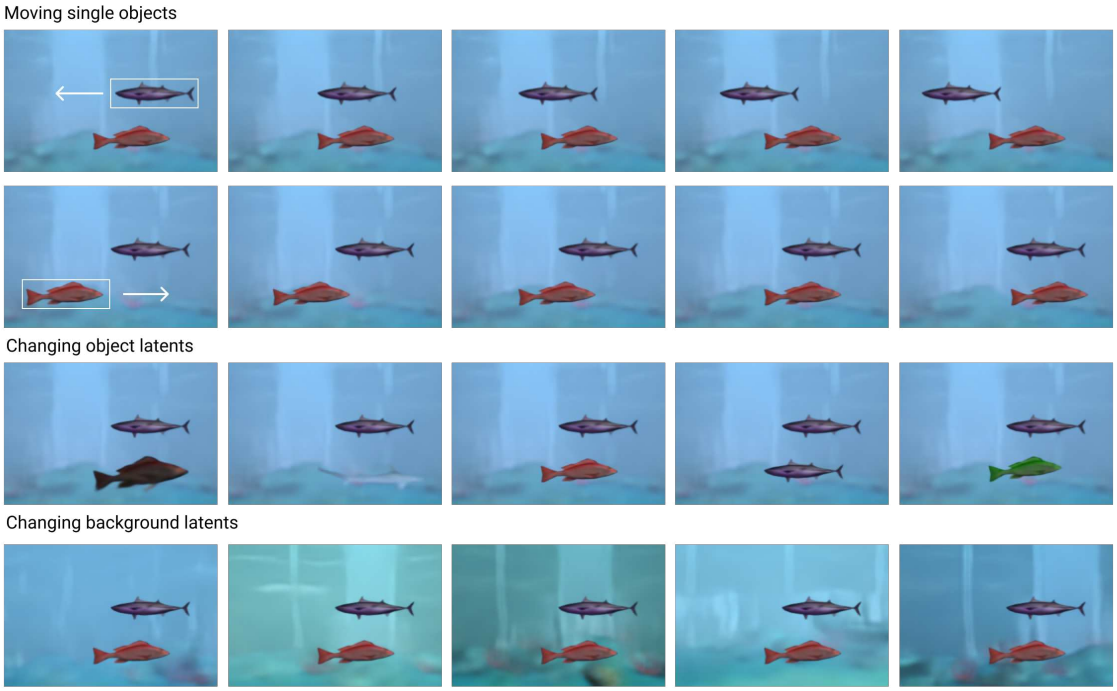

Figure 13: Two object "reconstructions". Object latents are obtained from a reference sample. Our modular scene model makes it straightforward to vary locations of single objects, exchanging single objects, or changing the background without affecting the output sample (top to bottom).

Figure 14: Additional samples from the scene model. Depicted samples use different (reconstructed) backgrounds, samples from the object model are obtained from the standard normal prior and filtered with 150 bit entropy threshold at $\tau = 0.2$, sample sizes are constrained on a reference training sample, object positions are sampled independently from a uniform prior. Samples are not cherrypicked.

C.6. Conditional sampling from the scene model

We present conditional samples from the scene model. As a simple baseline, we use a k-nearest neighbour approach for conditionally sampling object latents based on background latents. First, we extract a paired dataset of background and foreground latents from the training dataset. Second, we sample background latents from the standard Normal distribution (prior of the background model). Third, we compute the 2% nearest neighbouring videos based on the L^2 distance from the background latent. Fourth, we randomly sample a subset of foreground latents extracted by the motion segmentation. Finally, we reconstruct the scene using the background latents along with the chosen subset of foreground latents. We reject foreground latents with an entropy larger than 150 bit. Samples (non-cherrypicked) are depicted in Fig. 15.

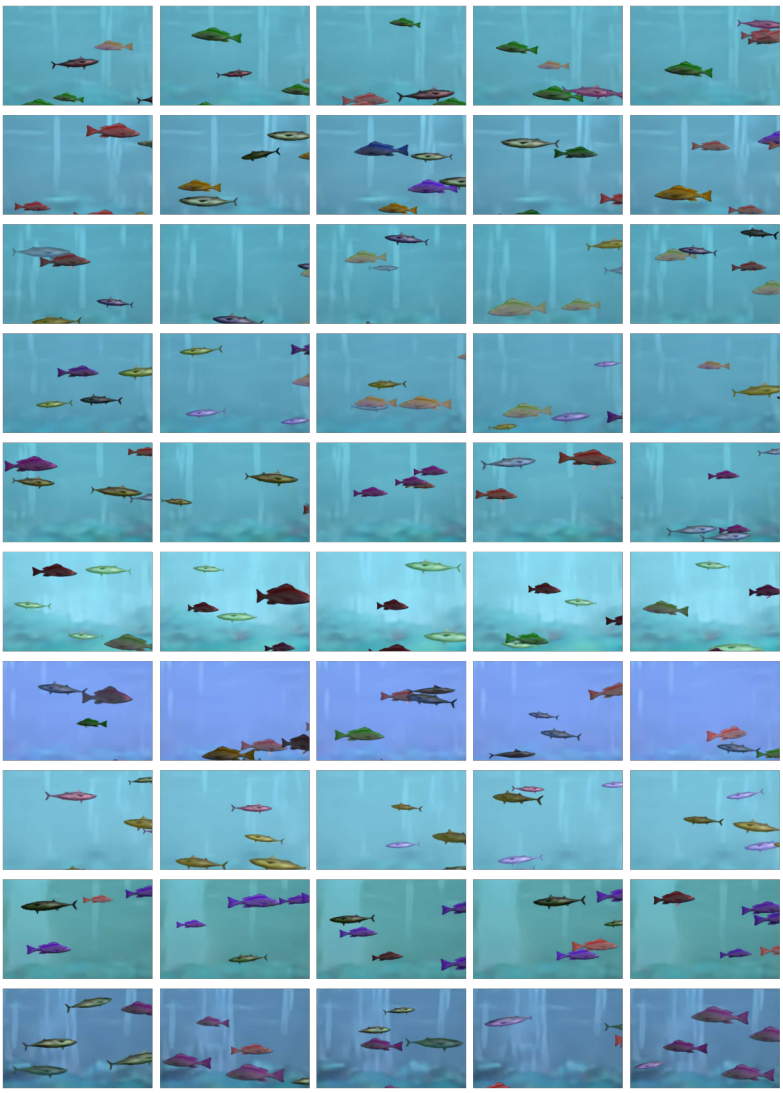

Figure 15: Conditional samples from the model. Fish appearances (qualitatively, mainly the brightness) now vary according to the background sample.

Similarly, we can constrain other latents like the x and y positions to a particular background. In Fig. 16 we show examples of constraining object locations based on a reference sample for both the ground truth and motion segmentation object model.

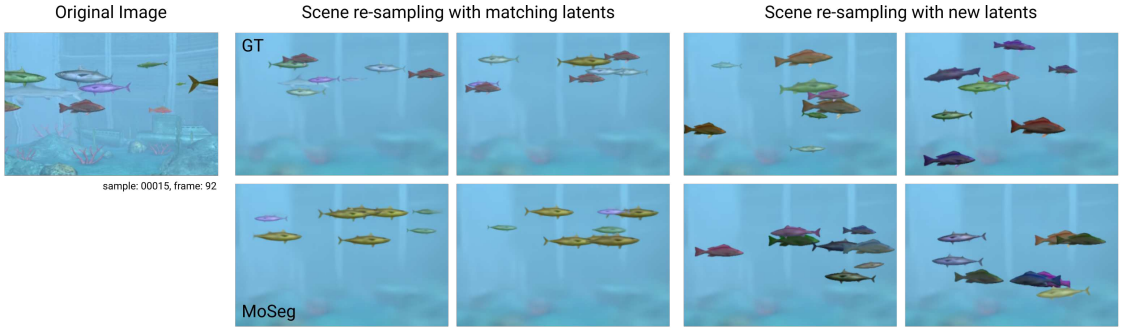

Figure 16: Conditional sampling of object locations based on a reference scene. For matching latents sampling, we extract the object latents and positions from the reference scene using the motion segmentation model. Samples obtained by matched sampling are more similar to the reference scene than samples obtained by unconstrained sampling from the model priors.

C.7. Entropy sampling

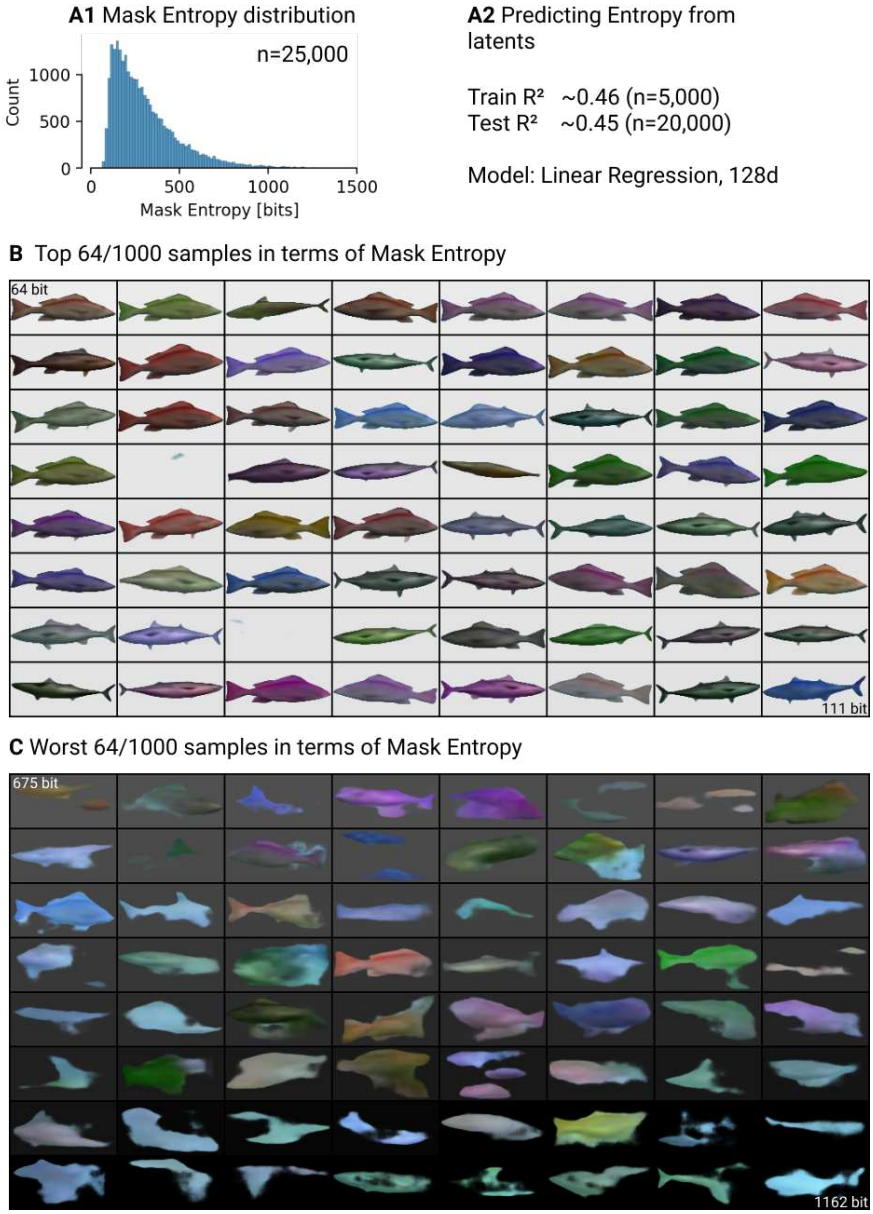

Figure 17: Details of entropy filtering when using the object model. (A1) Distribution of mask entropies for a given foreground model, estimated over 25,000 randomly sampled objects. The distribution typically peaks around 100 to 200 bits, makeing this a reasonable range for picking the cut-off. (A2) The 128d latent vector of objects is reasonably correlated with the entropy ($R^2 = 0.45$), making it possible to alter the prior to sample from to encourage low entropy samples. (B) 64 samples with the lowest entropy after drawing 1000 random samples from the object models and sorting according to entropy. While this strategy encourages some non-plausible, low entropy objects (row 4 and 6), it generally filters the dataset for samples with sharp boundaries and good visual quality. (C) 64 samples with the highest entropy after drawing 1000 random samples from the object models and sorting according to entropy. With a few exceptions of plausible samples (e.g. in row 2), most of the samples should be rejected in the scene model.

Appendix D. Comparison to related methods

D.1. GAN Baseline

As a baseline method for generating novel scenes for the FISHBOWL dataset, we consider the GAN used by Mescheder et al. (2018) for the CelebA-HQ dataset (Karras et al., 2018). We used the official implementation available at https://github.com/LMescheder/GAN_stability, and only changed the resolution of the generated images to 192×128px. For training, we used every 16th frame from each input video in the training set rescaled to the resolution of the GAN, resulting in 160k training images. We trained the model for 90 epochs.

Fig. 18 shows samples from the model in comparison to training frames. Overall, the GAN is able to generate images of convincing quality resembling the training images well. A particular strength of the GAN in comparison to our method, is it's ability to generate backgrounds with many details—which we explicitly left for future work. Several fish generated by the GAN however look distorted, in agreement with previous works concluding that GANs are good at generating "stuff", but have difficulties generating objects (Bau et al., 2019). This becomes especially apparent when visualizing interpolations in the latent space between samples, as done in Fig. 19.

Many differences between the samples generated by the GAN and our method, respectively, stem from the GAN being able to learn the overall statistics of the scenes well, but not learning a notion of objects. Compared to the GAN baseline, our object-centric approach offers several conceptual advantages: (1) The background and the objects in each scene are represented individually by our method, which makes it straightforward to intervene on the generated samples in a meaningful way (Fig. 6). While directions in the latent space of the GAN that correspond to semantically meaningful changes in the image might well exist, the GAN framework does not offer a principled way to find those directions without supervision. (2) While the GAN is able to generate novel scenes, it cannot be used to infer the composition of input scenes. (3) An optimally trained GAN perfectly captures the statistics of the input scene—making it impossible to generate samples beyond the training distribution. As our model explicitly represents scene parameters such as the number and position of fish, our model can be used for controlled out-of-distribution sampling (Fig. 6).

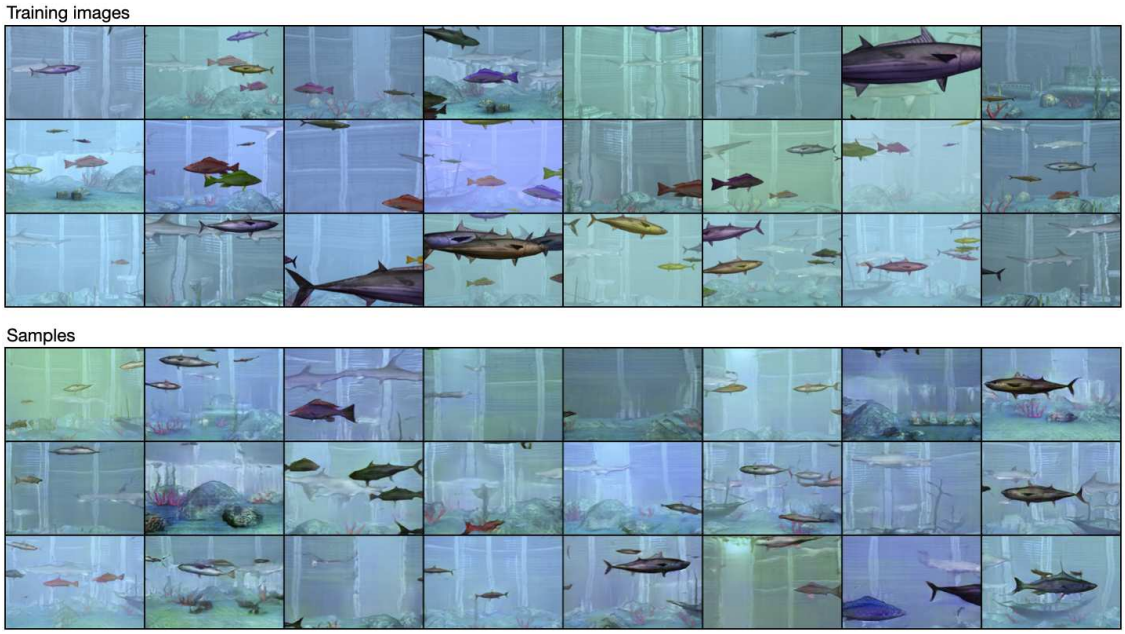

Figure 18: Samples from a baseline GAN (Mescheder et al., 2018) trained on frames from the training set of the FISHBOWL dataset. *Top:* Input frames from the dataset. *Bottom:* Samples generated by the GAN.

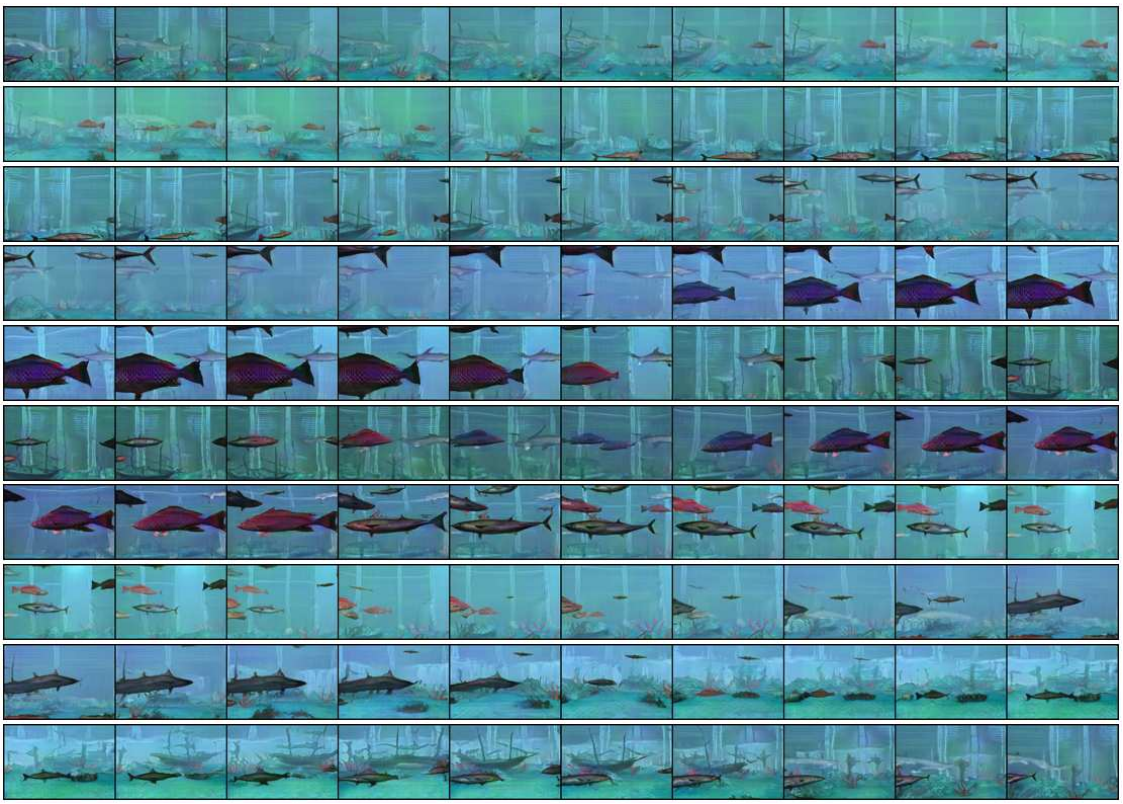

Figure 19: Samples from the baseline GAN, obtained by interpolating the latent vectors between random samples.

D.2. SPACE (Lin et al., 2020b)

SPACE (Lin et al., 2020b) is an object-centric representation learning method following the Attend-Infer-Repeat (AIR) framework (Eslami et al., 2016; Crawford and Pineau, 2019). Different from the AIR model, the detection of objects and inference of their latent representation is fully parallelized to make the approach faster and more scalable. We trained SPACE on the Fishbowl dataset using the implementation provided by the authors (https://github.com/zhixuan-lin/SPACE). We used the default hyperparameters and trained two variants using a 4x4 and 8x8 grid of object detectors, respectively. As for the GAN baseline in appendix D, we used every 16th frame for training this model to keep the training time reasonable (160k frames in total). Despite the subsampling, this is still substantially more data than the model was trained on in the original paper (60k images). SPACE expects the input images to have equal width and height, therefore we used a central square crop from every frame.

Training the model on a single Nvidia RTX 2080ti GPU is relatively fast (14h for the 4x4 grid and 18h for the 8x8 grid), confirming the performance advantage of the parallel object detectors. As the results in the Fig. 20 show, the object reconstructions from the model overall look reasonable. Most structure in the background is missed by the model, however we conjecture that this might be solvable by adapting the capacity of the background network. The visualization of the object detections however reveals a more fundamental failure mode due to the grid structure used by the

object detector in SPACE: Larger fish are split across several cells, even when using only 4x4 cells. As each cell can only handle at most one object, decreasing the cell count further is not expected to yield sensible results, as this limits the number of objects too much. Lin et al. (2020b) mentioned this problem and introduced a boundary loss to address it, however on the Fishbowl dataset this does not resolve the problem. We hypothesize that an approach based on a fixed grid, while adequate in some cases and having substantial performance advantages, doesn't work well with objects showing large scale variations as present in our newly proposed dataset. We believe this is a limitation of SPACE that cannot be resolved with further model tuning.

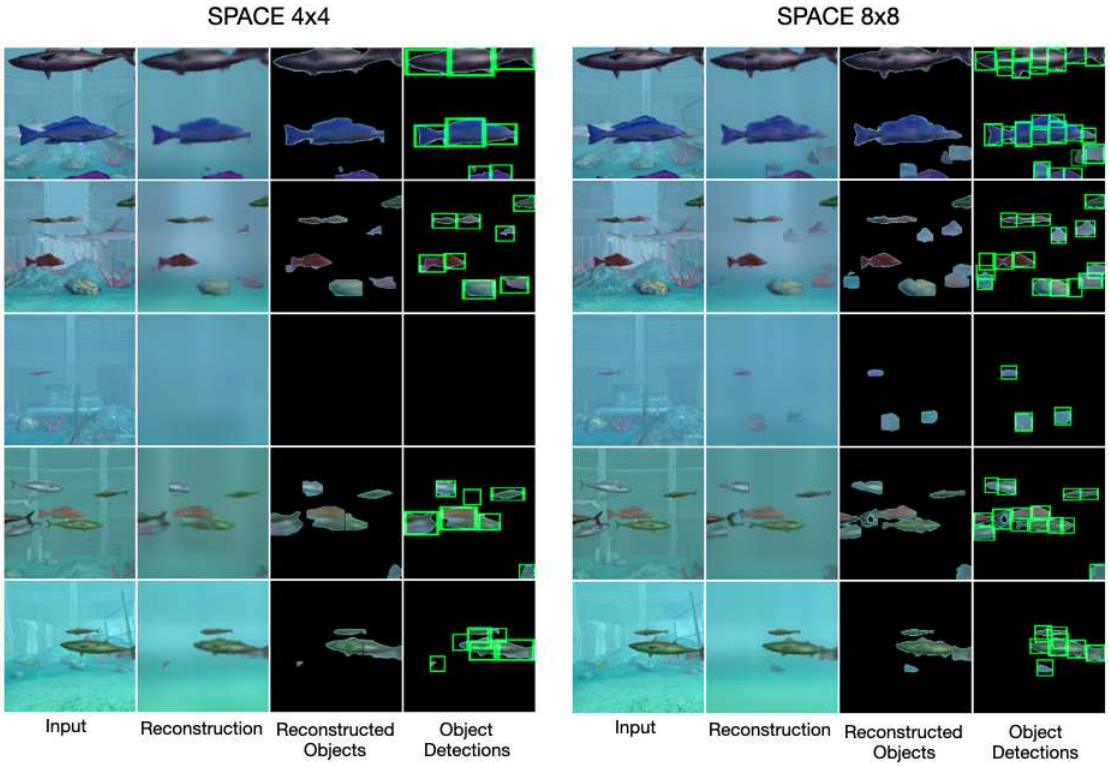

Figure 20: Scenes from the Fishbowl dataset reconstructed by SPACE Lin et al. (2020b). The model is trained using the official implementation provided by the authors with the default parameters. The two variants shown use grid sizes of 4x4 and 8x8, respectively.

We performed an additional experiment that integrates our approach with SPACE: As a first step, we trained a variant of our object model that uses the glimpse encoder and decoder from SPACE. We increased the learning rate to 0.0005 following an initial grid search, but otherwise left all hyperparameters unchanged. As the second step, we included the pretrained glimpse encoder and decoder into the SPACE model and trained the full model end-to-end as described above (using a grid size of 4x4). When we continued to train the glimpse encoder and decoder in the end-to-end setting, we noticed similar failures as described above. Therefore we froze the respective weights after the pretraining step.

Figure 21 shows results from this model variant. Early during the training, the model is in many cases able to detect the objects in the input scenes. This especially holds for larger objects that

have been shown to be problematic for the end-to-end model before. Later during training SPACE however increasingly uses the background model to also explain foreground objects. In addition, we evaluated both the pretrained and end-to-end trained object model from SPACE using the ground truth unoccluded objects. As the quantitative results in Table Tab. 4 show, the reconstruction of the appearance works better with the SPACE object model than with our architecture[6], but has a substantial gap in the segmentation performance. More importantly, the results clearly show an improvement of the motion segmentation based training over the original end-to-end approach, confirming the qualitative results.

Overall those results show that our motion segmentation based approach for object-centric modeling can also be used for scaling existing architectures to more complex settings. The off-the-shelf SPACE model however still falls short of our scene model when trained this way. In the future we see great potential in combining our object learning approach with state-of-the-art object-centric scene models.

6. To some degree this is confounded with the worse segmentation performance, as the appearance error is only evaluated on the intersection of the ground truth and predicted mask.

Table 4: Quantitative evaluation of the SPACE object model trained end-to-end within SPACE and pretrained based on motion segmentations using the same loss as our object model. The pretrained SPACE object model and ours use the cutout augmentation. The metrics are the same as in Tab. 2.

Object model	**IoU** ↑	**MAE** ↓	**IoU@0.5** ↑	**MAE@0.5** ↓
SPACE (end-to-end)	0.533	12.3	0.425	16.4
SPACE (pretrained)	0.734	**7.60**	0.605	**12.4**
Ours	**0.822**	13.0	**0.677**	24.0
Baseline	0.915		0.271	

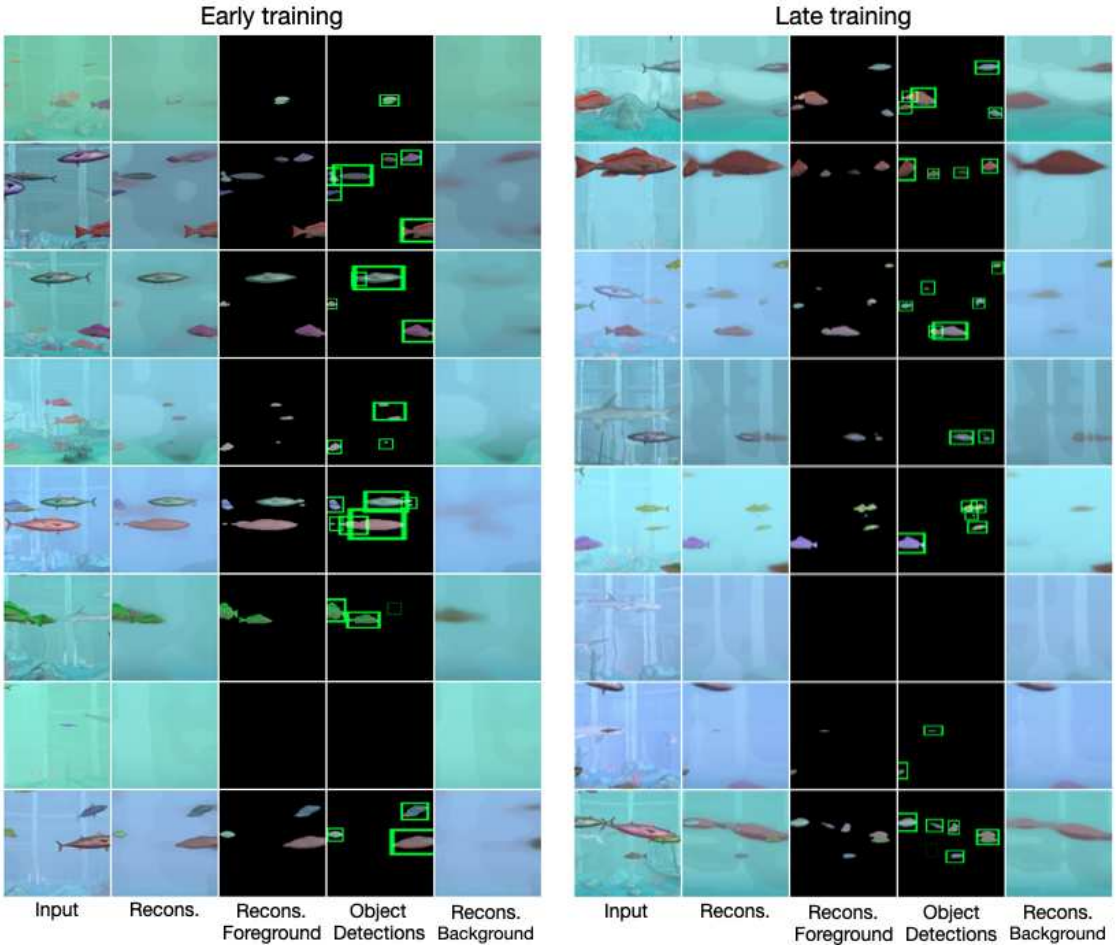

Figure 21: Scenes from the Fishbowl dataset reconstructed by SPACE (Lin et al., 2020b). The glimpse encoder and decoder where trained on the candidate objects given by the motion segmentation using the object model loss proposed in this work. Afterwards, the remaining components where trained end-to-end using the official implementation provided by the authors with the default parameters.

D.3. GENESIS-v2 (Engelcke et al., 2021)

We trained GENESIS-v2 on the Fishbowl dataset using the official implementation by the authors (https://github.com/applied-ai-lab/genesis) with default hyperparameters except for the image resolution, number of object slots and batch size. We modified the model code to work for rectangular images and used a resolution of 128x192 pixels. We trained GENESIS-v2 having 5 object slots on 4 Nvidia RTX 2080Ti GPUs using a batch size of 64. Initial experiments with 10 object slots lead to the background being split up into multiple slots. As for the GAN baseline before, we used every 16th frame for training this model (160k frames in total).

In Fig. 22 we show qualitative results from GENESIS-v2 on the Fishbowl dataset. The reconstructions of the model look somewhat blurry but capture the major structure in the input images well. Importantly, the visualization of the segmentation map and the individual slots reveal that the model succeeds in learning to decompose the scenes into background and objects. Sampling objects and scenes however fails with this model configuration. Most likely this happens due to the GECO objective (Rezende and Viola, 2018), that decreases the weight of the KL term in the loss function as long as the log likelihood of the input sample is below the target value. Training GENESIS-v2 with the original VAE objective instead of GECO leads to better scene samples, the decomposition of the scene however fails with this objective (Fig. 23).

As a comparison to the end-to-end training within GENESIS-v2, we trained a variant of our object model using the architecture of the GENESIS-v2 component decoder. As encoder, we use a CNN constructed symmetrically to the decoder, using regular convolutions instead of the transposed convolutions. We trained the model with the Cutout augmentation and using the same loss and training schedule as for the object model described in the main paper.

The results in Tab. 5 and Fig. 24 show that the model generally performs well, but worse than our original object model. This can most likely be explained by the larger capacity of our object model (10 vs 5 layers, 128 vs 64 latent dimensions) which seems to be necessary to model the visually more complex objects in our setting. Also when trained separately, the object model has difficulties with sampling novel fish, which can be addressed at the cost of worse reconstructions 25.

Overall we conclude that our modular object learning approach scales much better to visually more complex scenes than end-to-end training as done by GENESIS-v2. Even when using the GENESIS-v2 component decoder within our framework, the necessary trade-off between reconstruction and generation capabilities seems to be more favorable when using our modular approach as opposed to the end-to-end training. We remark that this comparison comes with a grain of salt: We neither adapted the hyperparameters of GENESIS-v2 nor of our method and the same decoder is used within GENESIS-v2 for the background, too. The strong qualitative difference in sample quality however makes it unlikely that this explains all of the difference. Moreover, our approach only addresses learning generative object models whereas GENESIS-v2 is also capable to infer the decomposition of static input scenes. For the future, we therefore see much potential in combining the respective strengths of both methods.

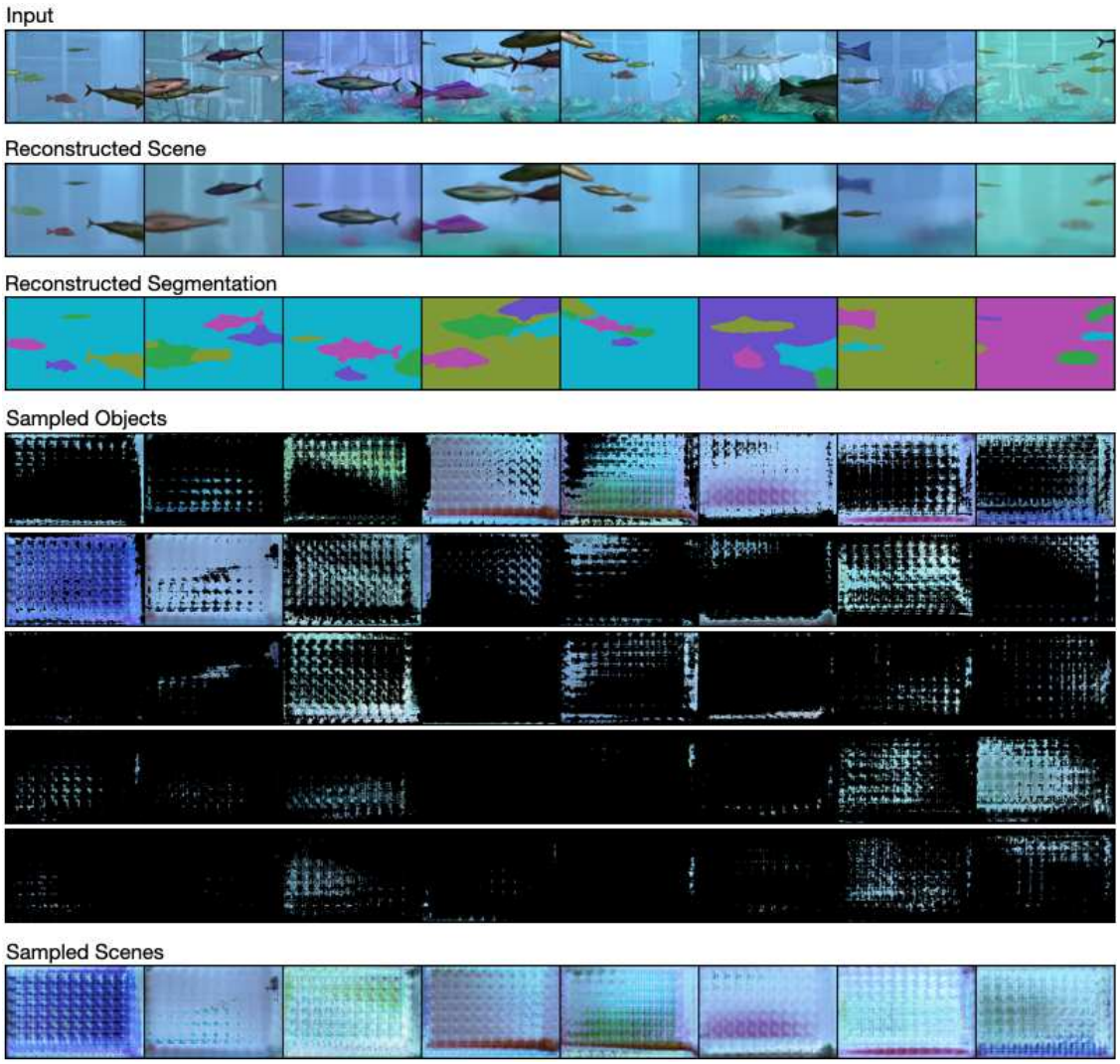

Figure 22: Qualitative results of GENESIS-v2 applied on the FISHBOWL dataset. The reconstruction in the second row look somewhat blurry, but capture all major structure in the input images shown in the first row. The visualization of the reconstructed segmentation shows that the model succeeds in decomposing the input image into the background and the different objects. Sampling from the model however fails in this setting.

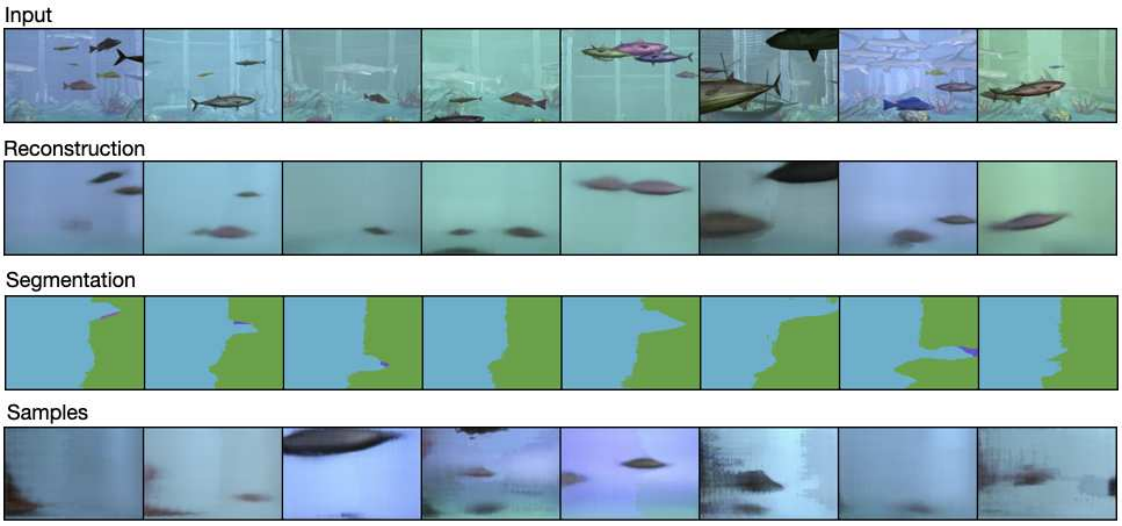

Figure 23: Qualitative results of GENESIS-v2 trained on the Fishbowl dataset using the default VAE objective instead of the GECO objective. Sampling from the model works much better in this setting; the model fails however to segment the scene into the indiviual objects.

Table 5: Comparison of the reconstructions from the object model based on the GENESIS-v2 component decoder with our original object model using the same metrics as in Tab. 2. Both models are trained using cutout augmentation.

Training data	**Architecture**	**IoU** ↑	**MAE** ↓	**IoU@0.5** ↑	**MAE@0.5** ↓
Motion Segmentation	GENESIS-v2	0.779±0.002	15.3±0.063	0.661±0.003	25.7±0.114
	Ours	0.822±0.001	13.0±0.032	0.677±0.002	24.0±0.017
Ground Truth Segmentation	GENESIS-v2	0.844±0.001	13.7±0.109	0.722±0.001	19.8±0.062
	Ours	0.887±0.001	12.3±0.035	0.743±0.003	17.3±0.087

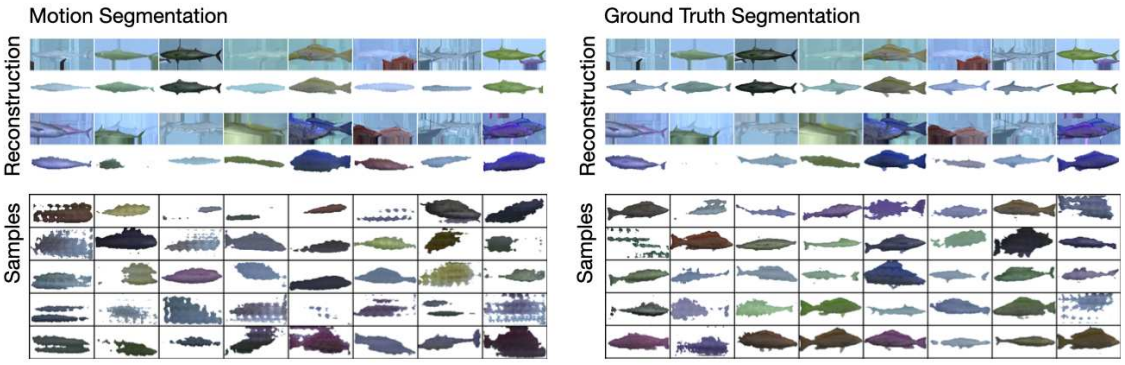

Figure 24: Qualitative results when using the GENESIS-v2 object decoder as object model within our modular training approach using the same loss and training schedule. Reconstructions and samples look worse than with the original object model, hinting at the larger capacity of our object model being necessary for our dataset.

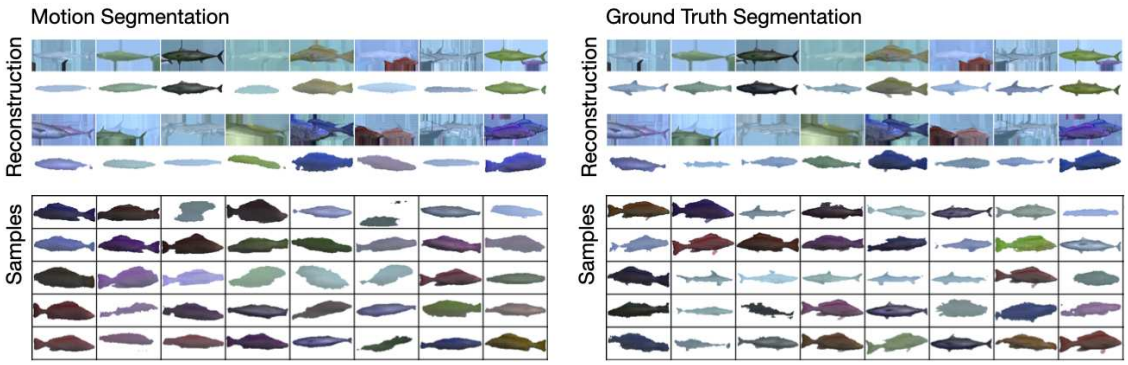

Figure 25: Qualitative results when using the GENESIS-v2 object decoder as object model within our modular training approach using a larger weight of the KL divergence in the VAE training loss. At the prize of worse reconstructions, the samples from the model can be substantially improved this way.

D.4. Comparison on the RealTraffic dataset

To evaluate how well our method transfers to other settings, we trained our model on the RealTraffic dataset (Ehrhardt et al., 2020). As the resolution of the images is smaller in the RealTraffic dataset, we reduced the object size threshold to 64px and the minimal distance to the boundary to 8px for the object extraction. We trained the object model with the same architecture as used for the Fishbowl dataset. Due to the smaller dataset size, we trained the model for 600 epochs and reduced the learning rate only after 400 epochs. As the videos in the dataset were all recorded from the same stationary traffic camera, we did not train a background model but used the mean-filtered background directly. In contrast to the aquarium dataset, the object positions and sizes are not distributed uniformly for this dataset. For the scene model, we therefore sample directly from the empirical joint distribution of object positions and sizes (extracted from the object masks obtained by the motion segmentation).

Figure 26: Results from our model trained on the RealTraffic dataset compared to results from other models trained on this dataset.

In Fig. 26, we compare samples from our model to other models for which results on the RealTraffic dataset have been reported by Ehrhardt et al. (2020). Overall, our model transfers well to this real world setting, given that our model is used largely unchanged. In comparison to the GAN-based RELATE and BlockGAN, the samples from our model look slightly more blurry. However the results from our method clearly improve over the VAE-based GENESIS model.

The adaptation of the scene statistics as described above works well for the object position, as the cars are correctly positioned on the street only. We consider the possibility of this straightforward adaptation to novel scene statistics to be a nice advantage of our object-centric modeling approach. In the future, this could be improved further by, e.g., learning to sample latent variables conditioned on the object positions.

Object segmentation from *common fate*: Motion energy processing enables human-like zero-shot generalization to random dot stimuli

Matthias Tangemann **Matthias Kümmerer** **Matthias Bethge**

University of Tübingen, Tübingen AI Center
matthias.{lastname}@bethgelab.org

Abstract

Humans excel at detecting and segmenting moving objects according to the *Gestalt* principle of "common fate". Remarkably, previous works have shown that human perception generalizes this principle in a zero-shot fashion to unseen textures or random dots. In this work, we seek to better understand the computational basis for this capability by evaluating a broad range of optical flow models and a neuroscience inspired motion energy model for zero-shot figure-ground segmentation of random dot stimuli. Specifically, we use the extensively validated motion energy model proposed by Simoncelli and Heeger in 1998 which is fitted to neural recordings in cortex area MT. We find that a cross section of 40 deep optical flow models trained on different datasets struggle to estimate motion patterns in random dot videos, resulting in poor figure-ground segmentation performance. Conversely, the neuroscience-inspired model significantly outperforms all optical flow models on this task. For a direct comparison to human perception, we conduct a psychophysical study using a shape identification task as a proxy to measure human segmentation performance. All state-of-the-art optical flow models fall short of human performance, but only the motion energy model matches human capability. This neuroscience-inspired model successfully addresses the lack of human-like zero-shot generalization to random dot stimuli in current computer vision models, and thus establishes a compelling link between the Gestalt psychology of human object perception and cortical motion processing in the brain.
Code, models and datasets are available at `https://github.com/mtangemann/motion_energy_segmentation`.

1 Introduction

Motion is a powerful cue that humans use to detect and segment visual objects. A striking example are camouflaged animals, which are difficult to spot when stationary but become much easier to detect when moving. Motion segmentation in humans is believed to be driven by the principle of common fate [47, 48, 42], which posits that elements that move together, belong together. Remarkably, human perception generalizes this principle in a zero-shot fashion to novel textures or moving random dots. For example, the seminal work by Johansson [19] showed that humans can easily detect biological motion from only few moving dots. More recently, Robert et al. [31] introduced random dot stimuli called *object kinematograms* that preserve the motion in a video while ensuring that static appearance cues are uninformative about the video contents (Fig. 1, example video in the supplemental material).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

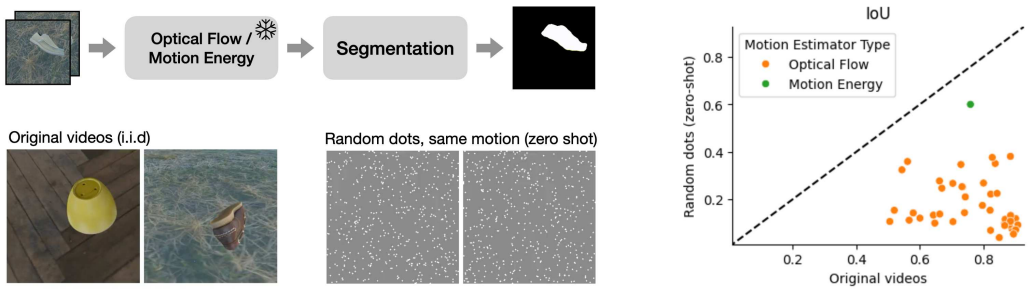

Figure 1: We compare state-of-the-art optical flow estimators and a neuroscience inspired motion energy model on a figure-ground segmentation task. For evaluation, we use random dot stimuli with the same motion patterns as the original videos, but for which the appearance of each individual frame is informative (example video in the supplemental material). The neuroscience inspired model generalizes to these stimuli much better than state-of-the-art optical flow models.

Nevertheless, humans were able to classify the animals and objects in these videos based on motion information alone.

In this work, we seek to understand the computational basis for appearance-agnostic motion perception in humans which enables this zero-shot generalization to random dots. Recent advancements in computer vision models for motion segmentation enable accurate segmentation of moving objects in natural videos based on a combination of optical flow estimation networks with downstream segmentation networks (e.g., [51]). However, it remains untested whether these models generalize in a similar way as human perception. Since the motion estimation stage is critical for segmenting moving objects, we focus on testing a broad range of state-of-the-art optical flow methods in combination with a fixed segmentation network. Our analysis reveals that existing computer vision approaches do not generalize in a human-like manner: Many high-performing models on natural stimuli perform near chance level for random dots.

In the primate visual cortex, area MT is known to be involved in motion perception and interpretation. Computational models for this area based on *motion energy* were proposed almost forty years ago [1, 46] and since then have been shown to predict key characteristics of neural firing patterns [37]. Instead of matching deep features between two frames, these models rely on spatio-temporal filtering in pixel space combined with a post-processing stage to resolve ambiguities. We demonstrate that this mechanism can be successfully integrated with deep neural networks for motion segmentation in realistic videos and reaches the performance of early deep learning based optical flow models on the original, textured videos—which is remarkable considering that the motion energy model was developed to explain the tuning of individual neurons and has several orders of magnitude fewer parameters than typical optical flow networks. Crucially, the motion energy model substantially outperforms all tested optical flow models in zero-shot generalization to moving random dots. In a direct comparison with humans in a controlled psychophysics study, the motion energy based approach is the only model that can match human capability.

In summary, our paper makes the following contributions:

- We show that a a broad range of state-of-the-art optical flow methods do not support human-like motion segmentation that generalizes to random dot patterns.
- We demonstrate that a classical neuroscience model can be successfully integrated with deep neural networks and generalizes to random dot stimuli.
- We conduct a psychophysical experiment to directly compare random dot motion segmentation in humans and machines. While state-of-the-art optical flow models fall short of human performance, the motion energy model can match it.

These results establish a compelling link between the Gestalt psychology of human object perception and cortical motion processing in the brain, showing that a motion energy approach can overcome the lack of human-like zero-shot generalization to random dot stimuli in current computer vision models. Integrating this mechanism with state-of-the-art optical flow methods is promising path towards more robust motion estimation models.

2 Related Work

Motion energy. Modelling motion perception in humans has been frequently approached using motion energy models. These models exploit the fact that a moving pattern corresponds to oriented edges when considering a video as a spatio-temporal volume [1, 46]. Several models have been proposed that build on this principle, aiming to explain the tuning properties of neurons found in visual areas V1 and MT [37, 15, 32, 28]. With few exceptions [41, 38], these models have not been used as a motion estimation models in a computer vision context. Our work is the first to study motion energy models for moving object segmentation.

Optical flow estimation. Optical flow traditionally has been formulated as an optimization problem with the goal of finding good matches between two frames [16]. During recent years, optimization based methods have been superseded by deep neural networks that frame optical flow estimation as an end-to-end regression task. FlowNet [11] pioneered this approach with a CNN that optionally includes an explicit temporal matching operation. Following works contributed better training data and proposed coarse-to-fine architectures to predict optical flow [17, 39, 40] which lead to substantial performance improvements. More recently, models that iteratively refine a high resolution optical flow map [43, 18] and Transformer-based models [36, 53, 54] have further improved state-of-the-art. Some works have compared optical flow models to human motion perception [56, 41], however not in the context of motion segmentation.

Motion Segmentation. The typical approach to motion segmentation is using optical flow as input for a downstream segmentation model. One classic line of work computes point trajectories from optical flow and then clusters the trajectories to segment moving regions [6, 29, 20]. Classical geometric approaches to motion segmentation have been combined with deep learning in later work [4, 5]. More recently, purely deep learning based approaches have been able to improve state-of-the-art [44, 9, 22, 23, 51]. To achieve high performance on classical motion segmentation datasets, the optical flow based motion segmentation is typically combined with appearance based segmentation [9, 52]. In this work, we evaluate generalization to random dot stimuli for which appearance is not informative, so we focus on purely motion-driven approaches.

3 Methods

The aim of this work is to evaluate which computational models match the capabilities of humans for zero-shot motion segmentation of random dot patterns. We follow the standard motion segmentation approach in computer vision and first use a motion model to estimate the motion in an input video, followed by a segmentation network that predicts the foreground mask. In order for models to perform well on zero-shot segmentation of random dot patterns, it is critical that the motion estimator used by the model generalizes well to these random dot stimuli. Ideally, the motion estimator would be invariant to changes in texture. Therefore, we focus on the motion estimation stage by evaluating a broad range of optical flow models in comparison to a neuroscience inspired motion energy model. As a segmentation model, we use the same segmentation architecture for all motion estimators which we train from scratch for every model.

3.1 Optical Flow Models

We use a range of optical flow models that includes all major deep learning based approaches to optical flow estimation. FlowNet 2.0 [17] was the first CNN based model that reached the performance of classical, optimization based methods. We consider three variants of the model using different combinations of subnetworks. PWC-Net [39] introduced a multi-scale approach that combined operations from classical approaches (such as cost volumes and warping), with components from deep learning. Different from previous models, RAFT [43] is not based on a coarse-to-fine approach but rather on iterative refinement of a high resolution optical flow map derived from multi-scale correspondences. GMA [18] extends the RAFT architecture by introducing a Transformer-based module to better handle occlusions, which have been shown to be difficult for previous models. More recently, GMFlow [53, 54] and FlowFormer++ [36] have been proposed as fully Transformer-based architectures for optical flow estimation.

We use the implementations and checkpoints of these models from the MMFlow library [8], except for FlowFormer++ and GMFlow for which we use the implementations and checkpoints provided by the respective authors[1][2]. For each architecture, we consider checkpoints trained on different datasets that are common in the field, including the FlyingChairs [11], FlyingThings3d [24], Sintel [7] and KITTI [25, 26]. In total, we evaluate 40 optical flow models.

We apply the models to predict multi-scale optical flow, in order to match the multi-scale features predicted by the motion energy model. All of the optical flow models internally use several scales to predict optical flow. However, this representation is followed by non-trivial processing to combine motion information across scales, so that using this internal representation directly would most likely lead to inferior performance. Therefore we use the unmodified models and scale the final optical flow prediction to the desired resolutions using bilinear interpolation.

3.2 Motion Energy Model

Motion energy models are based on the insight that a motion pattern in a video corresponds to a spatio-temporal orientation when the video is considered as an x-y-t volume [1, 46]. The motion at every pixel can therefore be estimated by using spatiotemporal filters that respond to a particular motion direction and speed. This mechanism has important differences from the optical flow models discussed before. All of the optical flow models compute deep features for two frames individually and match these features between two frames to estimate motion. The spatio-temporal filters in motion energy models on the other hand operate directly in pixel space. This approach leads to more ambiguous matches, which are typically resolved by considering more than two frames, and a postprocessing stage.

In this study, we build on the influential motion energy model by Simoncelli & Heeger [37]. In addition to the oriented filters described above, this model introduced a second stage that implements an *intersection of constraints* construction [13, 2] in order to resolve ambiguities of the linear filter responses. This motion energy model can be implemented as a CNN with the architecture shown in Figure 2. We derived the weights of the CNN from the the parameters of the original model and verified that our PyTorch [30] implementation of the motion energy model equals the original MATLAB implementation[3] up to numerical differences. Following the original model, we apply the model for five different input scales that are obtained by repeatedly blurring and downsampling the input by a factor of two. To streamline the implementation, we do not scale the activations after every layer and experimentally verified that this change does not affect downstream performance for motion segmentation.

3.3 Segmention model

We use a coarse-to-fine segmentation network to predict per-pixel logits for the respective pixel belonging to the foreground object (Figure 2). Input to the segmentation model are the multi-scale motion energy maps or multi-scale optical flow maps as predicted by the models described earlier. At each scale, the segmentation model consists of three components: The *input projection* layer predicts motion features for each scale. The core of the network is a *refinement CNN* that aggregates features across scales. At each scale, the refinement CNN concatenates the motion features from the current scale with the refined representation from all previous scales and predicts the refined representation for the current scale. Finally, the *output projection* layer predicts the segmentation given the refined representation from the finest scale. All layers except for the output projection are followed by a CELU nonlinearity [3] and instance normalization [45]. The parameters of the components are shared across the stages, so that the network is essentially a recurrent neural network that integrates information from coarsest to the finest scale in order to predict a segmentation.

3.4 Training

All models are trained on a synthetic video dataset that we generated using the Kubric library[14]. Each video shows a single moving object in front of a moving background. The 3D objects and

[1] `https://github.com/XiaoyuShi97/FlowFormerPlusPlus`

[2] `https://github.com/autonomousvision/unimatch`

[3] `https://www.cns.nyu.edu/~lcv/MTmodel/`

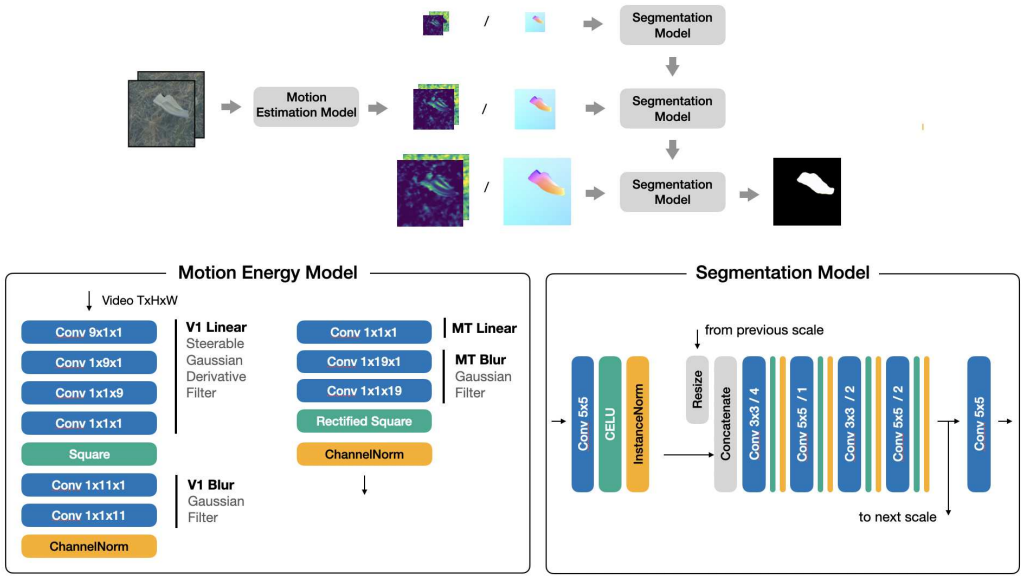

Figure 2: (*top*) Our motion segmentation architecture: The motion estimation predicts multi-scale optical flow or motion energy, the segmentation model predicts the moving foreground region. (*bottom left*) The motion energy model is implemented as a CNN. The weights are chosen such that the CNN is equivalent to the original model by [37]. (*bottom right*) The segmentation model combines motion features across scale and predicts a binary segmentation at the input resolution.

backgrounds used for dataset generation are scans of everyday objects and scenes, resulting in highly realistic renderings. We used 901 videos for training and 100 test videos, each having 90 frames at 30Hz. The training and test videos used different sets of object and backgrounds but are otherwise sampled from the same distribution. The code and hyperparameters for generating videos, as well as the rendered dataset, are publicly available[4].

For all models, we freeze the weights of the motion estimator and only train the downstream segmentation network. As common for binary motion segmentation, we use per pixel binary cross entropy to the ground truth masks as loss. We use the Adam optimizer [21] with a learning rate of $1e-4$ for all models and train for 40.000 steps using a batch size of 8. All models are trained on NVIDIA GeForce RTX 2080 Ti GPUs with 12GB of VRAM. Depending on the computational requirements of the motion model, training the segmentation model on a single GPU takes between 2 and 6 hours.

3.5 Zero-shot evaluation on random dot stimuli

We evaluate models on the original test videos as well as random dot stimuli generated for all test videos based on the ground truth optical flow. We use the same procedure as [31] for generating random dot stimuli using 500 dots with a lifetime of 8 frames, which matches the dot density and lifetimes.

We apply all models using a shifting window approach for the full length videos, but excluding the first and last four frames so that the window is fully contained within the video for all models. For evaluation, we obtain a binary prediction by thresholding with 0.5 and measure performance by computing IoU and F-Score for each frame individually and then averaging over the test set.

[4] `https://github.com/mtangemann/motion_energy_segmentation`

4 Results

4.1 Zero-Shot Random Dot Segmentation

Table 1 summarizes the motion segmentation performances achieved when using different motion estimators, both on the i.i.d. test set and the corresponding random dot stimuli. For a better overview, we visualize the performances as measured by IoU in Figure 1 and in the appendix.

		Original		Random Dots	
Motion Estimator	Training Dataset	IoU	F-Score	IoU	F-Score
Motion Energy (ours)	-	0.759	0.845	**0.600**	**0.718**
GMFlow	Flying Things 3D	0.885	0.925	0.381	0.493
(2 scales, 6 refinements)	Sintel	0.823	0.874	0.222	0.315
	Mixed	0.823	0.874	0.069	0.106
FlowNet2 SD	FlyingChairs	0.828	0.884	0.377	0.499
FlowNet2 CSS	FlyingThings3D	0.837	0.896	0.351	0.469
	FlyingChairs	0.735	0.818	0.252	0.359
PWC-Net	FlyingThings3D	0.729	0.815	0.347	0.469
	FlyingChairs	0.544	0.662	0.324	0.442
	KITTI	0.600	0.715	0.121	0.193
FlowNet2	FlyingChairs	0.662	0.757	0.278	0.380
	FlyingThings3D	0.821	0.877	0.154	0.241
FlowNet2 CS	FlyingThings3D	0.800	0.867	0.269	0.374
	FlyingChairs	0.669	0.760	0.247	0.359
GMFlow (2 scales)	Flying Things 3D	0.704	0.793	0.267	0.373
	Sintel	0.733	0.812	0.253	0.351
	Mixed	0.743	0.822	0.210	0.300
GMFlow (1 scale)	Mixed	0.843	0.897	0.225	0.308
	Flying Things 3D	0.797	0.859	0.174	0.244
GMA (+P)	KITTI	0.520	0.630	0.154	0.224
	FlyingChairs	0.646	0.717	0.101	0.142
	Mixed	0.895	0.932	0.054	0.078
	FlyingThings3D	0.850	0.894	0.039	0.057
FlowFormer++	Flying Chairs	0.741	0.800	0.144	0.218
	Flying Things 3D	0.901	0.935	0.119	0.182
	Sintel	**0.908**	**0.942**	0.092	0.140
	Flying Things 3D	0.902	0.938	0.072	0.113
GMA	KITTI	0.579	0.679	0.143	0.214
	FlyingChairs	0.643	0.724	0.134	0.194
	FlyingThings3D	0.867	0.909	0.100	0.150
	FlyingThings3D + Sintel	0.890	0.928	0.089	0.132
	Mixed	0.890	0.927	0.077	0.109
GMA (P-only)	FlyingChairs	0.663	0.743	0.138	0.207
	KITTI	0.567	0.650	0.112	0.159
	Mixed	0.881	0.920	0.093	0.141
	FlyingThings3D	0.867	0.909	0.090	0.142
RAFT	FlyingThings3D + Sintel	0.886	0.924	0.132	0.180
	FlyingThings3D	0.869	0.909	0.116	0.172
	Mixed	0.885	0.925	0.108	0.145
	KITTI	0.506	0.600	0.107	0.147
	FlyingChairs	0.647	0.718	0.100	0.147

Table 1: Model performances for the i.i.d. test videos and zero shot to the corresponding random dot stimuli with the same motion patterns. For all motion estimators, the same segmentation network is used to predict the figure-ground segmentation. Results are grouped by the motion estimation model and ordered by the performance on the random dot stimuli.

Recent optical flow methods perform strongly on the original videos. FlowFormer++ works best on our dataset with an IoU of 90.8%, closely followed by a GMA variant that reaches 89.5% IoU.

These results parallel the strong performance of recent Transformer-based architectures on standard optical flow benchmarks. The motion energy based model only achieves a performance of 75.9% IoU and lags behind state-of-the-art optical flow models, but performs similar as earlier deep learning based optical flow models. This result is remarkable when considering that the motion energy model predates the deep learning models by several decades and has not been tuned for dense, end-to-end motion prediction. Within each model, the checkpoints from the FlyingThings3d dataset tend to perform best for the original videos. The FlyingThings3d dataset contains renderings of 3D objects undergoing rigid motion, so arguably it is the most similar dataset compared to the one used in this study.

Motion energy generalizes much better to random dots. The motion energy based model reaches an IoU of 60.0%, which outperforms the performance of the second best model by more than 20 percentage points. Strikingly, the FlowFormer++ and GMA models that performed best on the original videos generalize particularly bad to the random dot stimli (IoU < 10%). Overall, more dated optical flow architectures such as FlowNet2 variants and PWC-Net tend to generalize better to random dot stimuli than more recent approaches. An interesting exception is GMFlow, which reached an IoU of 38.1% and performed best among all optical flow models. We do not observe a clear effect of the training dataset.

We visualize model predictions in Figure 3. For the original videos, the quality of the predicted optical flow varies but allows for a clear segmentation of the moving object. The object is also clearly represented in the motion energy maps, with some feature maps responding highly to the background and others to the moving object. The motion energy maps however tend to be noisier than the optical flow predictions, which explains the lower performance of the motion energy model for the clean videos.

The random dot stimuli exhibits the same motion as the original video, so the prediction of an ideal motion estimator would be unchanged. The optical flow methods however fail to properly estimate the motion of the foreground object. While some methods like FlowNet 2.0 and PWC-Net predict a highly noisy motion pattern that roughly matches the location of the foreground object, many optical flow estimators fail to detect the foreground motion at all. The motion energy on the other hand looks highly similar for the random dot stimulus and the original video, allowing the motion energy segmentation to generalize well in this case.

4.2 Ablation study

As an ablation study, we evaluated whether the performance of the motion energy segmentation model can be improved by learning the parameters of the motion model. We tested different combinations of layers in the motion energy CNN that are fixed, finetuned or trained from scratch and trained them end-to-end with the segmentation model.

The results in table 2 show that the original weights of the model allow for the best generalization to random dots. This is remarkable when considering that the weights of the motion energy model have been originally selected to explain the tuning properties of individual neurons, but not for image-computable motion estimation. Some of the configurations however outperformed the original weights on the original videos. So while the network architecture allows for generalization in principle, all our models trained by gradient descent converged to solutions that performed well on the training data but did not generalize.

As a further ablation study, we removed or replaced layers of the motion energy model. The results in the supplemental information suggest that the pooling and normalization layers are particularly important for generalization to random dots. More details and further experiments are provided in the supplemental information.

5 Human Machine Comparison

The previous results have revealed differences between different motion estimation models in terms of generalization to random dots. While it is known that humans can recognize objects in random dot stimuli without prior training [31], the ability to segment objects in moving random dot patterns has not been quantified before. We therefore conduct a human subject study in order to directly compare the zero-shot generalization to random dots in humans and machines.

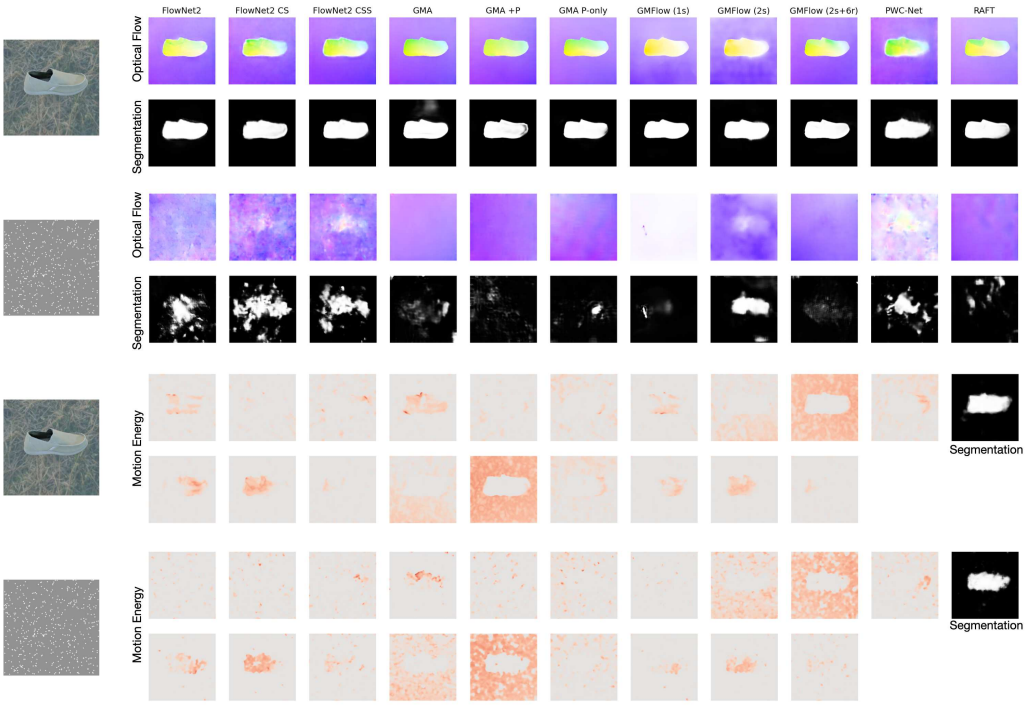

Figure 3: Example predictions for different motion estimators. The motion pattern in the random dot stimulus is the same as in the original video. While the optical flow estimates are highly accurate for the original videos, the models struggle with the random dot stimuli that exhibit the same motion. The activations of the motion energy model model however generalize well to the random dot stimuli, enabling to detect and segment the foreground object.

				Original		Random Dots	
V1 Linear	V1 Blur	MT Linear	MT Blur	IoU	F-Score	IoU	F-Score
fix	fix	fix	fix	0.759	0.845	**0.600**	**0.718**
fix	fix	finetune	fix	0.776	0.856	0.563	0.686
finetune	fix	finetune	fix	0.804	0.873	0.468	0.599
fix	fix	scratch	scratch	0.794	0.873	0.463	0.583
finetune	fix	scratch	scratch	**0.827**	**0.887**	0.395	0.508
finetune	finetune	scratch	cratch	0.600	0.717	0.162	0.246
scratch	fix	scratch	fix	0.660	0.752	0.052	0.087
scratch	scratch	scratch	scratch	0.593	0.702	0.027	0.048

Table 2: Comparison of using the original weights (fix), finetuning the original weights (finetung) or training from scratch (scratch) for the layers of the motion energy model.

Due to the inherent difficulty in directly evaluating the segmentation perceived by humans, we employed a shape identification task as a surrogate requiring segmentation (Figure 4). Each trial involved a random target shape and a distractor shape from the Infinite dSprites dataset [12]. The random dot stimulus shows the target shape moving linearly across the center of the image with random motion direction and speed. After the video concluded, participants were shown clean renderings of the target and distractor shapes and were required to select the shape that they perceived in the random dot stimulus. Since the shape alternatives were unknown while the random dot stimulus was shown, participants had to segment and memorize the shape in the random dot video and then compare it to the shape choices afterward. Therefore, performing well on this task necessitates sufficiently good segmentation of the moving shapes within the motion patterns of the random dot stimuli.

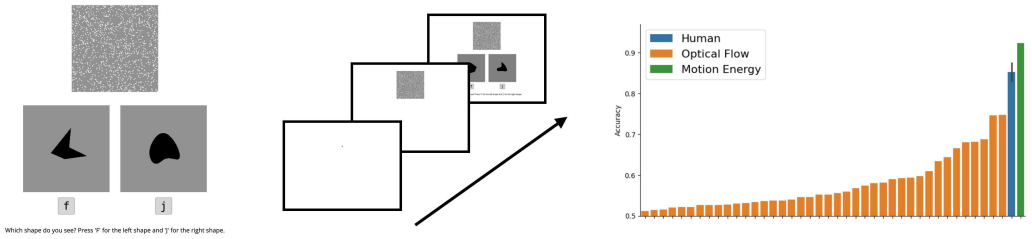

Figure 4: We compare humans and machines using a *random dot shape identification* task as a proxy to measure segmentation in humans. Shown a video of random dots, participants have to respond which of two shapes was perceived in the video. Humans outperformed all optical flow based models, but not the motion energy based model for this task. More details are provided in the supplemental material.

We performed the study in a controlled vision lab environment, where participants viewed the experiment on a VIEWPixx 3D LCD monitor (1920x1080, 120Hz) with the distance fixed to 65cm using a chin rest. The duration of all videos was 1s at a framerate of 30 Hz. Overall, we collected data from N=13 subjects, of which we excluded one subject due to insufficient visual acuity (remaining: N=12, 4 female, 8 male). Among the subjects where both trained vision scientists and naive subjects.

We evaluated all models on the same stimuli as human subjects. Given a random dot video, we applied the respective model to segment the video and selected the shape option that better matched the prediction as measured by IoU.

The results in Figure 4 show that all models based on optical flow are clearly outperformed by humans. Many of the optical flow based models perform near chance level, while some models reached a non-trivial performance. Overall, more recent optical flow models that perform very well on the original videos appear to generalize worse to this task, with GMFlow [53, 54] being a noteable exception. Different from the optical flow models, the motion energy based approach is the only model to match and even outperform human performance. More detailed results in the supplemental information show that the motion energy segmentation model performs on par with the highest performing participants of the study.

6 Limitations

To allow for comparing a large number of motion estimation models with a reasonable computational budget we made compromises for other modeling aspects. We limited the size of the segmentation network to allow for efficient training but performed a control experiment to show that using a more sophisticated segmentation network does not improve generalization (see supplemental information). Moreover, we used the same training schedule for all models but ensured that our setting supports all models adequately by visually inspecting the loss curves.

When comparing humans and machines we did not model several factors that are expected to influence human performance, such as the impact of internal noise and attentional lapses. As common in psychophysics experiments, several subjects reported making accidental errors for few examples [50] which negatively affects performance. So even a model that perfectly replicates the motion processing algorithm in humans is not expected to perfectly replicate human behavior in our setting.

7 Broader impact

This work is highly interdisciplinary, bridging state-of-the-art computer vision motion segmentation algorithms with the principles of Gestalt psychology and the neuroscience of cortical motion processing in the brain. By showing that a neuroscience-inspired motion energy model can outperform conventional optical flow models in zero-shot generalization to random dot stimuli, the study highlights the potential for integrating biological insights into AI systems. Benefits of broader impact include the development of more robust and human-like AI systems, educational value, and the creation of AI systems that are more aligned with human cognition.

8 Discussion

Computational models for motion estimation have a long history in both computational neuroscience and computer vision. Shallow models based on spatio-temporal filtering in pixel space have been able to predict neural activity in brain areas related to motion perception [37, 32] and are compatible with a range of phenomena in human perception [57]. In computer vision, models based on matching deep features between two frames have continuously improved performance over the last years and are successfully applied in a range of downstream tasks. Despite these successes, our study reveals a striking gap between deep optical flow networks and human perception: While humans generalize the common fate Gestalt principle to zero-shot segmentation of random moving dots, the optical flow models fail to generalize to these stimuli. Furthermore, we show that a classic motion energy approach can be scaled to realistic videos while matching human generalization capabilities.

The great success of deep neural networks in computer vision has spawned interest in using DNNs also as a model for human vision, in particular for core object recognition [55]. In the same spirit, deep neural networks might be promising models for human motion perception [56]. While promising, our study parallels findings for core object recognition that show striking differences between human perception and DNNs [49]. For motion perception, however, we show that is possible to combine classical models from computational neuroscience with the scalability of deep learning. Further integration of these modeling traditions is a promising path towards image computable models of human motion perception [41].

While closing the gap between human perception and machine vision is crucial for computational neuroscience, we believe that computer vision likely profits from better alignment with human vision as well. Humans still greatly outperform machines in terms of robustness and efficiency. Our study suggests a substantial entanglement of motion estimation with appearance in DNNs, which might also be linked to the lack of robustness observed in state-of-the-art motion estimators [33, 34]. Computational principles that better match human vision should be considered as promising candidates for addressing these issues.

Finally, we argue that deep learning based models as presented in our work have the potential to greatly improve our understanding of motion perception in humans. Low-level mechanisms for motion estimation and higher level processes for motion interpretation have been mostly studied in isolation [27]. In our work we follow a more holistic approach by studying the effects motion detection mechanisms on the perception of moving objects, which offers several unique opportunities. First, it is not necessary for most downstream tasks to perfectly estimate the physically correct motion. For example, segmenting moving objects does require precise information about object boundaries while other mistakes are less critical. Studying motion estimation and interpretation jointly allows to better understand viable compromises in estimation accuracy as the basis for more efficient processing. Second, studying end-to-end models of motion estimation and interpretation advances our understanding of how neural mechanisms give rise to behavior. DNNs are a particularly promising modeling approach positioned in a "Goldilocks zone" regarding the trade-off between biological plausibility and scalability to natural stimuli and tasks [10]. In this vein, our work establishes a compelling link between cortical mechanisms for motion estimation and the Gestalt psychology of human object perception.

In the future, this work can be extended in several directions. While scaling remarkably well, the original motion energy model is not able to match the performance of state-of-the-art optical flow methods on natural scenes. We see integrating principles from computational neuroscience with techniques from deep learning as a promising path towards closing this gap [41]. Moreover, training the parameters of the CNN implementation of the motion energy model jointly with the segmentation model did not lead to a generalizable solution. How humans learn generalizable motion perception from data, or to which degree this capability is innate, are important questions for future research. Finally, in the spirit of the neuroconnectionist research programme [10] we see our model as an executable hypothesis for motion perception in the human brain. While matching human performance in terms of generalization to moving random dots, this model might well fail to capture other aspects of human motion perception. Further evaluating and extending models of motion perception to capture a diverse range of phenomena is an exciting path towards a holistic understanding of human perception.

Acknowledgments and Disclosure of Funding

This work was supported by the German Research Foundation (DFG): SFB 1233, Robust Vision: Inference Principles and Neural Mechanisms, TP 4, project number: 276693517. The authors thank the International Max Planck Research School for Intelligent Systems for supporting MT. We thank Felix Wichmann, Thomas Klein and all other members of the Wichmann-Lab for supporting and testing the human machine comparison study, and Larissa Höfling for valuable feedback on the manuscript.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, February 1985. doi: 10.1364/JOSAA.2.000284.

[2] E. H. Adelson and J. A. Movshon. Phenomenal coherence of moving visual patterns. *Nature*, 300(5892):523–525, December 1982. doi: 10.1038/300523a0.

[3] J. T. Barron. Continuously Differentiable Exponential Linear Units. *arXiv preprint arXiv:1704.07483*, April 2017. doi: 10.48550/arXiv.1704.07483.

[4] P. Bideau and E. Learned-Miller. It's Moving! A Probabilistic Model for Causal Motion Segmentation in Moving Camera Videos. In *Computer Vision – ECCV 2016*, pages 433–449, Cham, October 2016. Springer International Publishing. doi: 10.1007/978-3-319-46484-8_26.

[5] P. Bideau, A. RoyChowdhury, R. R. Menon, and E. Learned-Miller. The Best of Both Worlds: Combining CNNs and Geometric Constraints for Hierarchical Motion Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 508–517, June 2018.

[6] T. Brox and J. Malik. Object Segmentation by Long Term Analysis of Point Trajectories. In *Computer Vision – ECCV 2010*, volume 6315 of *Lecture Notes in Computer Science*, pages 282–295, Berlin, Heidelberg, September 2010. Springer. doi: 10.1007/978-3-642-15555-0_21.

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A Naturalistic Open Source Movie for Optical Flow Evaluation. In *Computer Vision – ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 611–625, Berlin, Heidelberg, October 2012. Springer. doi: 10.1007/978-3-642-33783-3_44.

[8] M. Contributors. MMFlow: OpenMMLab Optical Flow Toolbox and Benchmark, 2021.

[9] A. Dave, P. Tokmakov, and D. Ramanan. Towards Segmenting Anything That Moves. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2019.

[10] A. Doerig, R. P. Sommers, K. Seeliger, B. Richards, J. Ismael, G. W. Lindsay, K. P. Kording, T. Konkle, M. A. J. van Gerven, N. Kriegeskorte, and T. C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, July 2023. doi: 10.1038/s41583-023-00705-w.

[11] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning Optical Flow With Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, December 2015.

[12] S. Dziadzio, Ç. Yıldız, G. M. van de Ven, T. Trzciński, T. Tuytelaars, and M. Bethge. Infinite dSprites for Disentangled Continual Learning: Separating Memory Edits from Generalization. In *3rd Conference on Lifelong Learning Agents (CoLLAs)*, July 2024.

[13] C. L. Fennema and W. B. Thompson. Velocity determination in scenes containing several moving objects. *Computer Graphics and Image Processing*, 9(4):301–315, April 1979. doi: 10.1016/0146-664X(79)90097-2.

[14] K. Greff, F. Belletti, L. Beyer, C. Doersch, Y. Du, D. Duckworth, D. J. Fleet, D. Gnanapragasam, F. Golemo, C. Herrmann, T. Kipf, A. Kundu, D. Lagun, I. Laradji, H.-T. D. Liu, H. Meyer, Y. Miao, D. Nowrouzezahrai, C. Oztireli, E. Pot, N. Radwan, D. Rebain, S. Sabour, M. S. M. Sajjadi, M. Sela, V. Sitzmann, A. Stone, D. Sun, S. Vora, Z. Wang, T. Wu, K. M. Yi, F. Zhong, and A. Tagliasacchi. Kubric: A Scalable Dataset Generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3761, June 2022.

[15] D. J. Heeger, E. P. Simoncelli, and J. A. Movshon. Computational models of cortical visual processing. *Proceedings of the National Academy of Sciences*, 93(2):623–627, January 1996. doi: 10.1073/pnas.93.2.623.

[16] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1): 185–203, August 1981. doi: 10.1016/0004-3702(81)90024-2.

[17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of Optical Flow Estimation With Deep Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, July 2017.

[18] S. Jiang, D. Campbell, Y. Lu, H. Li, and R. Hartley. Learning To Estimate Hidden Motions With Global Motion Aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9772–9781, October 2021.

[19] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, June 1973. doi: 10.3758/BF03212378.

[20] M. Keuper, B. Andres, and T. Brox. Motion Trajectory Segmentation via Minimum Cost Multicuts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3271–3279, December 2015.

[21] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations (ICLR) 2015*. arXiv, May 2015. doi: 10.48550/arXiv.1412.6980.

[22] H. Lamdouar, C. Yang, W. Xie, and A. Zisserman. Betrayed by Motion: Camouflaged Object Discovery via Motion Segmentation. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[23] H. Lamdouar, W. Xie, and A. Zisserman. Segmenting Invisible Moving Objects. In *32nd British Machine Vision Conference (BMVC)*, November 2021.

[24] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, June 2016.

[25] M. Menze, C. Heipke, and A. Geiger. JOINT 3D ESTIMATION OF VEHICLES AND SCENE FLOW. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II-3/W5:427–434, September 2015. doi: 10.5194/isprsannals-II-3-W5-427-2015.

[26] M. Menze, C. Heipke, and A. Geiger. Object Scene Flow. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140:60–76, June 2018. doi: 10.1016/j.isprsjprs.2017.09.013.

[27] S. Nishida, T. Kawabe, M. Sawayama, and T. Fukiage. Motion Perception: From Detection to Interpretation. *Annual Review of Vision Science*, 4(1):501–523, September 2018. doi: 10.1146/annurev-vision-091517-034328.

[28] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. *Current Biology*, 21(19): 1641–1646, October 2011. doi: 10.1016/j.cub.2011.08.031.

[29] P. Ochs and T. Brox. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *2011 International Conference on Computer Vision*, pages 1583–1590, November 2011. doi: 10.1109/ICCV.2011.6126418.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., December 2019.

[31] S. Robert, L. G. Ungerleider, and M. Vaziri-Pashkam. Disentangling Object Category Representations Driven by Dynamic and Static Visual Input. *Journal of Neuroscience*, 43(4):621–634, January 2023. doi: 10.1523/JNEUROSCI.0371-22.2022.

[32] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon. How MT cells analyze the motion of visual patterns. *Nature Neuroscience*, 9(11):1421–1431, November 2006. doi: 10.1038/nn1786.

[33] J. Schmalfuss, L. Mehl, and A. Bruhn. Attacking Motion Estimation with Adversarial Snow. In *ECCV 2022 Workshop on Adversarial Robustness in the Real World*, October 2022.

[34] J. Schmalfuss, P. Scholze, and A. Bruhn. A Perturbation-Constrained Adversarial Attack for Evaluating the Robustness of Optical Flow. In *Computer Vision – ECCV 2022*, pages 183–200, Cham, October 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-20047-2_11.

[35] H. H. Schütt, S. Harmeling, J. H. Macke, and F. A. Wichmann. Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, 122: 105–123, May 2016. doi: 10.1016/j.visres.2016.02.002.

[36] X. Shi, Z. Huang, D. Li, M. Zhang, K. C. Cheung, S. See, H. Qin, J. Dai, and H. Li. FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1599–1610, June 2023.

[37] E. P. Simoncelli and D. J. Heeger. A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761, March 1998. doi: 10.1016/S0042-6989(97)00183-1.

[38] F. Solari, M. Chessa, N. V. K. Medathati, and P. Kornprobst. What can we expect from a V1-MT feedforward architecture for optical flow estimation? *Signal Processing: Image Communication*, 39:342–354, November 2015. doi: 10.1016/j.image.2015.04.006.

[39] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8934–8943, June 2018.

[40] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz. Models Matter, So Does Training: An Empirical Study of CNNs for Optical Flow Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(6):1408–1423, June 2020. doi: 10.1109/TPAMI.2019.2894353.

[41] Z. Sun, Y.-J. Chen, Y.-H. Yang, and S. Nishida. Modeling Human Visual Motion Processing with Trainable Motion Energy Sensing and a Self-attention Network. In *Advances in Neural Information Processing Systems*, volume 36, pages 24335–24348. Curran Associates, Inc., December 2023.

[42] M. Tangemann, S. Schneider, J. von Kügelgen, F. Locatello, P. V. Gehler, T. Brox, M. Kümmerer, M. Bethge, and B. Schölkopf. Unsupervised Object Learning via Common Fate. In *Proceedings of the Second Conference on Causal Learning and Reasoning*, pages 281–327. PMLR, April 2023.

[43] Z. Teed and J. Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *Computer Vision – ECCV 2020*, pages 402–419, Cham, August 2020. Springer International Publishing. doi: 10.1007/978-3-030-58536-5_24.

[44] P. Tokmakov, C. Schmid, and K. Alahari. Learning to Segment Moving Objects. *International Journal of Computer Vision*, 127(3):282–301, March 2019. doi: 10.1007/s11263-018-1122-2.

[45] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv preprint arXiv:1607.08022*, November 2017. doi: 10.48550/arXiv.1607.08022.

[46] A. B. Watson and A. J. Ahumada. Model of human visual-motion sensing. *Journal of the Optical Society of America A*, 2(2):322–342, February 1985. doi: 10.1364/JOSAA.2.000322.

[47] M. Wertheimer. Experimentelle Studien über das Sehen von Bewegung. *Zeitschrift für Psychologie*, 61:161–265, April 1912.

[48] M. Wertheimer. *On Perceived Motion and Figural Organization*. The MIT Press, Cambridge, MA, July 2012.

[49] F. A. Wichmann and R. Geirhos. Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, 9:501–524, September 2023. doi: 10.1146/annurev-vision-120522-031739.

[50] F. A. Wichmann and N. J. Hill. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, 63(8):1293–1313, November 2001. doi: 10.3758/BF03194544.

[51] J. Xie, W. Xie, and A. Zisserman. Segmenting Moving Objects via an Object-Centric Layered Representation. In *Advances in Neural Information Processing Systems*, volume 35, pages 28023–28036. Curran Associates, Inc., December 2022.

[52] J. Xie, W. Xie, and A. Zisserman. Appearance-Based Refinement for Object-Centric Motion Segmentation. *arXiv preprint arXiv:2312.11463*, August 2024. doi: 10.48550/arXiv.2312.11463.

[53] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao. GMFlow: Learning Optical Flow via Global Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8121–8130, June 2022.

[54] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger. Unifying Flow, Stereo and Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (11):13941–13958, November 2023. doi: 10.1109/TPAMI.2023.3298645.

[55] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, May 2014. doi: 10.1073/pnas.1403112111.

[56] Y.-H. Yang, T. Fukiage, Z. Sun, and S. Nishida. Psychophysical measurement of perceived motion flow of naturalistic scenes. *iScience*, 26(12), December 2023. doi: 10.1016/j.isci.2023.108307.

[57] J. Yates. Motion Illusions, December 2020. `https://jake.vision/blog/motion-illusions/` (accessed: 2024-10-24).

Supplemental information

A Additional details about the results

For an additional overview, view visualize the segmentation performances on random dot stimuli as reported in Table 1.

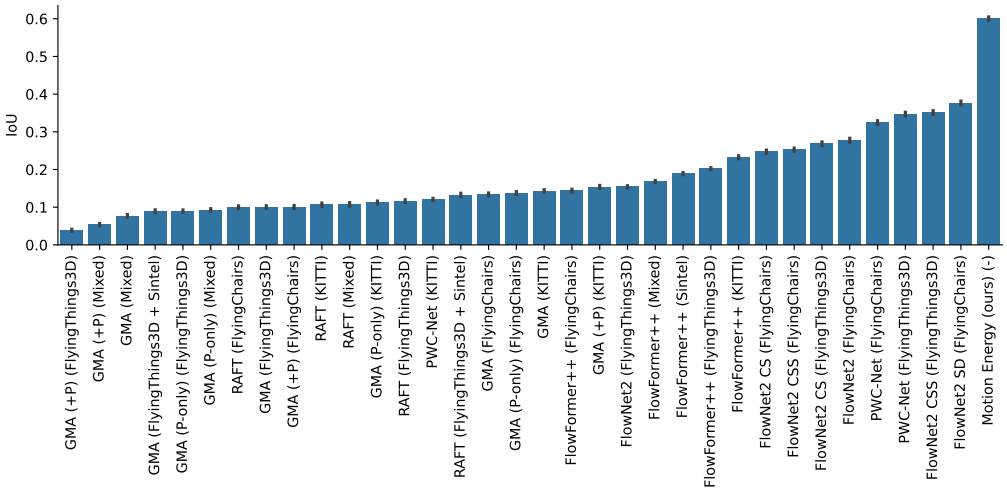

Figure 5: Segmentation performances of the evaluated models on the random dot stimuli. Same data as in Table 1.

B Additional experiments

B.1 Importance of components of the motion energy model

We conducted an additional ablation study in order to better understand which aspects of the motion energy model are essential for generalization to random dot stimuli. We removed or replaced individual layers as described in Table 3 and trained the ablated models from scratch using in the same way as the baseline model.

The results in Table 2 hint at the normalization and pooling layers being important for generalization. When the Gaussian pooling layers are removed completely, the performance on original videos even slightly improves while the generalization to random dot stimuli is substantially reduced.

Replacing the squaring-based nonlinear layers with ReLU layers, however, hardly changes the model's performance.

Condition	Original		Random Dots	
	IoU ↑	F-Score ↑	IoU ↑	F-Score ↑
Baseline	0.759	0.845	0.600	0.718
Replace RectifiedSquare → ReLU (MT)	0.753	0.838	**0.609**	**0.725**
Replace Square → ReLU (V1)	0.770	0.854	0.536	0.663
Remove MT Linear	0.768	0.856	0.481	0.609
Remove MT	0.770	0.854	0.451	0.583
Remove Blur (V1, MT)	**0.801**	**0.872**	0.421	0.540
Replace ChannelNorm → InstanceNorm (V1, MT)	0.592	0.703	0.230	0.340
Remove Normalization (V1, MT)	0.400	0.516	0.018	0.018

Table 3: Ablation study: Performance of the model on original videos and corresponding random dot stimuli with various layers of the motion energy model removed or replaced. Results are ordered by IoU on the random dot stimuli.

B.2 Multi-frame optical flow

The motion energy model uses a window of 9 frames as input, while typical optical flow methods estimate correspondences between only two frames. To rule out the possibility that the results observed in our paper are mainly explained by the different input window lengths, we perform an ablation study in which we apply optical flow methods using the same 9 frame windows. For each window, we compute the optical flow between the central frame, for which the segmentation has to be predicted, to the 8 other frames in the window. The stacked optical flow fields are then used as the input to the segmentation network.

The results in Table 4 and Figure 6 show some improvement on the original videos but an ever wider gap to the motion energy model in terms of of generalization to random dots. The differences between the motion energy and optical flow models therefore cannot be explained by the different input lengths.

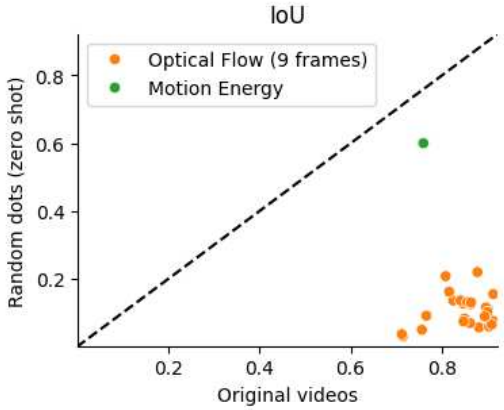

Figure 6: Performance of multi-frame optical flow based models on the original videos and corresponding random dot videos.

B.3 Comparison with state-of-the-art motion segmentation

In our study we used a relatively small segmentation network downstream to the respective motion estimator. State-of-the-art motion segmentation models typically target multi-object segmentation in real world videos and therefore use more complex segmentation networks. In order to verify that the results in our paper are not caused by using a smaller segmentation network, we evaluated the state

Motion Estimator	Training Dataset	Original IoU	Original F-Score	Random Dots IoU	Random Dots F-Score
Motion Energy (ours)	-	0.759	0.845	**0.600**	**0.718**
FlowNet2 SD	FlyingChairs	0.878	0.928	0.221	0.325
FlowNet2	FlyingChairs	0.808	0.868	0.209	0.300
	FlyingThings3D	0.881	0.929	0.058	0.100
PWC-Net	FlyingChairs	0.816	0.886	0.163	0.250
	FlyingThings3D	0.825	0.886	0.137	0.221
	KITTI	0.712	0.811	0.038	0.060
RAFT	FlyingThings3D + Sintel	**0.912**	**0.948**	0.156	0.222
	FlyingChairs	0.863	0.914	0.126	0.195
	Mixed	0.896	0.934	0.117	0.164
	FlyingThings3D	0.894	0.934	0.090	0.132
	KITTI	0.714	0.794	0.031	0.053
FlowNet2 CS	FlyingChairs	0.841	0.899	0.137	0.220
	FlyingThings3D	0.847	0.904	0.075	0.129
GMA (+P)	FlyingChairs	0.856	0.912	0.132	0.212
	Mixed	0.900	0.936	0.114	0.179
	FlyingThings3D	0.899	0.936	0.104	0.171
GMA	FlyingChairs	0.864	0.917	0.131	0.212
	Mixed	0.900	0.937	0.090	0.139
	FlyingThings3D + Sintel	0.909	0.943	0.066	0.100
	FlyingThings3D	0.903	0.943	0.060	0.098
	KITTI	0.756	0.834	0.051	0.084
GMA (P-only)	FlyingChairs	0.846	0.901	0.128	0.207
	KITTI	0.766	0.847	0.092	0.155
	FlyingThings3D	0.903	0.940	0.083	0.139
	Mixed	**0.912**	0.947	0.077	0.117
FlowNet2 CSS	FlyingChairs	0.850	0.908	0.084	0.141
	FlyingThings3D	0.862	0.918	0.070	0.121

Table 4: Ablation study: We apply the optical flow estimators to a window of 9 frames by using the central frame as references and computing optical flow to each of the 8 other frames. The stacked optical flow fields are used as inpute for the segmentation network.

of the art OCLR model [51] in our setting. The OCLR model uses optical flow estimated by RAFT [43], which we also included in our experiments. The segmentation network however uses a U-Net architecture with Transformer bottleneck and was trained to segment multiple objects on a synthetic dataset. We use the published weights and do not retrain the model on our data.

The results in Table 5 show that the model performs very well on the original data. OCLR outperforms our motion energy based model and achieves a performance similar to the best optical flow based models considered in this work. At the same time, the model does not generalize to the corresponding random dot stimuli. These results provide further evidence that the low generalization to random dots is not due to the architecture of the segmentation network or the RGB training data, but a property of the motion estimator.

Model	**IoU (original)**	**IoU (random dots)**
OCLR	0.838	0.026
Motion Energy Segmentation	0.759	0.600

Table 5: Comparison of the state-of-the-art motion segmentation model OCLR, and our segmentation model based on a motion energy model.

C Additional details about the human subject study

C.1 Comparison of humans and machines by example difficulty

As a measure of task difficulty, we count the number of *informative dots*. A dot is informative, if it is contained in either the target and distractor shape but not both (see Figure 7, left). Only these dots allow discriminating between the different shapes.

We fitted psychometric curves for human participants and models as a function of the number of informative dots, using the psignifit toolbox [35]. The results in Figure 7 confirm that only the motion energy model is able to match the performance of human subjects, especially for stimuli with a medium number of informative dots.

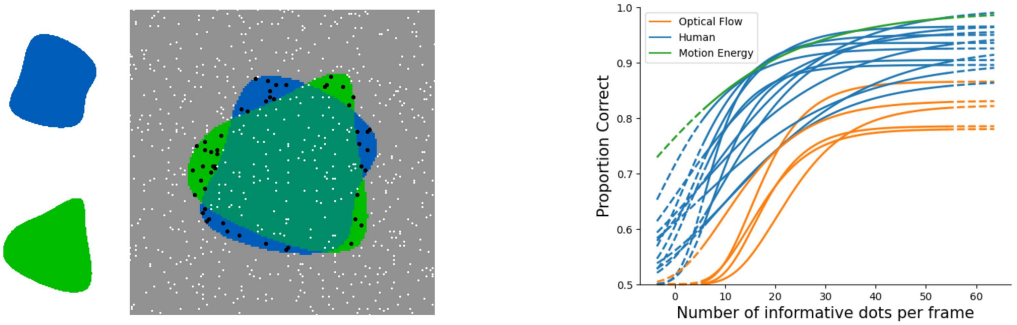

Figure 7: (*left*) As a measure of task difficulty, we count the number of informative dots that allow discriminating betwen the two shape alternatives. (*right*) Psychometric curves for humans, the motion energy based model and the four best optical flow models for the task as in 8.

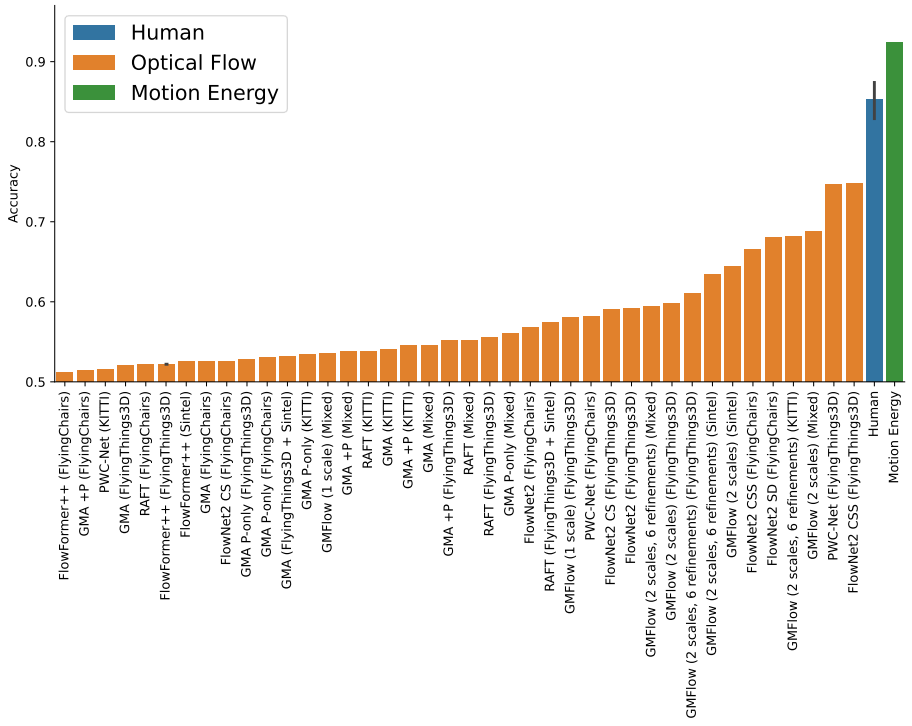

Figure 8: Comparison of the human and model performances for the random dot shape matching task.

C.2 Screenshots of the experiment

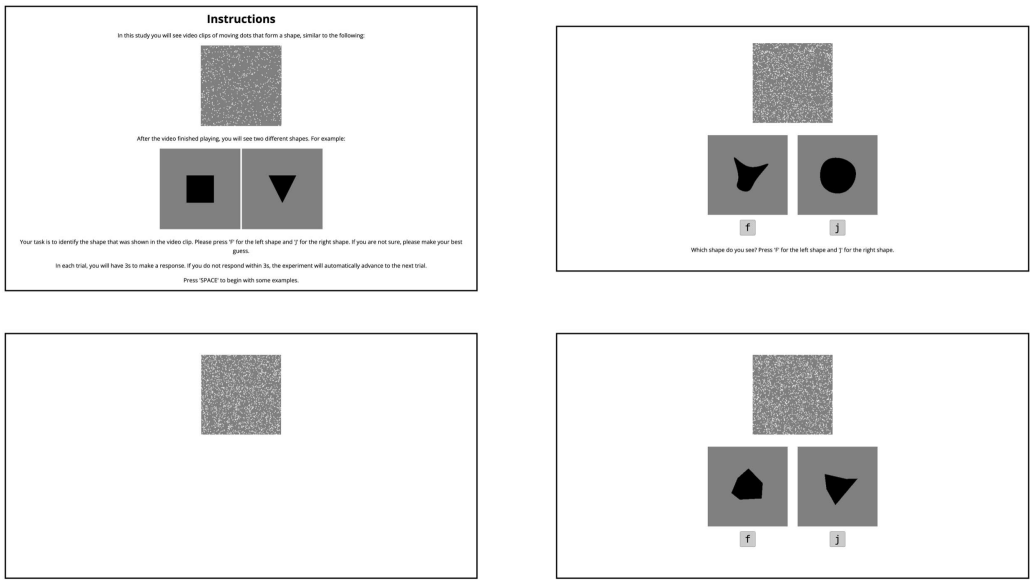

Figure 9: Screenshots from the human subject study on random dot shape identification. (*top left*) Instructions that were shown prior to the experiment. (*top right*) We showed 20 training trials during which subjects could familiarize themselves with the task. (*bottom left*) The training was followed by 500 test trials. A video with the random dot stimuli was shown first. (*bottom right*) Once the video finished playing, the two shape options were shown below.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the main contributions of our work in both the abstract and the introduction. All mentioned results are supported by the experimental data presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a separate section that discusses the limitations of our work in detail, including limitations due to computational constraints.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the models, data and evaluation protocol used in the paper in detail. Additionally, the code, pretrained models and the contributed dataset are publicly released.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All code and data for the paper is publicly released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Hyperparameters and training details are explicitly reported with the description of the models.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Due to space constraints we did not include further statistical information in the main table. However we included a Figure showing the same data with error bars in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We reported the computational resources of our models with the description of the training details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the code of ethics and conform to it in every respect.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We are discussing potential broader impacts of our work in a dedicated section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper does not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used the Kubric generator with the built-in asset library and a range of pretrained models. We cited the sources of all data and implementations that we used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The synthetic data we generated for the paper is described in the paper, and published with the code used to generate it.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Screenshots of the experiment are included in the supplemental material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The study in this paper does not pose any particular risk on participants. IRB approval exists.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.