Department of Physics and Astronomy

University of Heidelberg

Master thesis

in Physics

submitted by

Leonard Bereska

born in Celle

2019

Unsupervised Disentanglement of Geometric Shape and Visual Appearance

This Master thesis has been carried out by Leonard Bereska

at the

Heidelberg Collaboratory for Image Processing

under the supervision of

Prof. Björn Ommer

Unüberwachtes Trennen von geometrischer Form und visueller Erscheinung

Objektrepräsentationen sind von fundamentaler Bedeutung für viele Anwendungen in Computer Vision. Die vorliegende Arbeit präsentiert und evaluiert einen unüberwachten Ansatz, um eine kompositionelle Teilrepräsentation von Objekten zu lernen, welche die geometrische Form und die visuelle Erscheinung für jeden Objektteil trennt. Um die Faktoren von Form und Erscheinung zu trennen, wird Invarianz unter Transformationen des jeweils anderen Faktors angenommen. Zusätzlich wird verwendet, dass räumliche Form equivariant bezüglich räumlicher Transformationen ist. Diese Annahmen werden in eine Autoencoder Architektur eingebaut, die so konstruiert ist, dass eine lokale Interpretation der Objektteile gewahrt bleibt.
Die Methode weist dem Objekt Teile zu, die das Objekt konsistent und sinnvoll abdecken, ohne manuelle Überwachung oder a priori Annahmen über die Objektklasse zu benötigen. Die Methode wird evaluiert auf einer Anzahl an herausfordernden Datensätzen, wobei die Objektklassen variieren: von menschlichen Gesichtern über menschliche Personen bis hin zu Hunden, Katzen und Vögeln. Das unüberwachte Lernen von Form wird evaluiert, indem aus den entdeckten Teilen, Objektmarkierungen regressiert werden. Der neueste Stand der Forschung wird hierbei signifikant geschlagen. Darüber hinaus wird gezeigt, dass die Repräsentation tatsächlich in Form und Erscheinung aufspaltet und lokale Teile unabhängig voneinander lernt.

Unsupervised Disentanglement of Geometric Shape and Visual Appearance

Object representations are of paramount importance for various computer vision applications. The thesis at hand presents an unsupervised approach to learn a dedicated compositional part representation of an object, disentangling the factors of shape and appearance for each part. To disentangle, each factor is assumed to be invariant under transformations of the other factor. Additionally, shape is assumed to be equivariant with respect to a spatial transformation. These assumptions are implemented in a two-stream auto-encoding framework for detecting parts, while the architecture is designed to maintain the local nature of the parts.
Trained without any manual supervision or prior information on the object class, the method discovers parts consistently, covering the whole object. We evaluate this on a diverse selection of challenging datasets, the object classes ranging from human faces and bodies to dogs, cats and birds. The model outperforms state-of-the-art methods significantly in terms of unsupervised landmark regression. Additionally, we show that our model actually learns to disentangle shape from appearance and learns local parts independently.

Contents

1 Introduction

Computer vision is the scientific endeavour to algorithmically understand patterns in images. Structures and objects in the physical world interact in complex ways to generate an image. The image then acts as a mirror, in which these elements of the world are reflected and leave patterns. To recognize patterns in an image, means in essence to observe the reality lurking behind this mirror, *i.e*. to measure the causal elements that contributed to the image generation.

1.1 Why Disentangle Generative Factors?

Figure 1.1: *"Imagine, how ridiculous you would look if you wore that hot pants"* - Thought experiments are a targeted manipulation of a disentangled representation.

On the one hand, there are pragmatic reasons to aim at extracting disentangled factors from images: to successfully transfer a representation between different tasks, typically only a few factors are relevant [1]. Efficient transfer and multi-task learning should account for this. On the other hand, learning to capture external mechanisms in appropriate internal representations, can be seen as a step to automate visual reasoning itself. Once disentangled, a factor can be manipulated individually to make a targeted change in an image. Thereby, not only humans may change images at will, but also machines may reason about the world [2], by simulating changes to factors internally in their model of the world. Thought experiments like *"imagine, how ridiculous you would look, if you wore that hot pants"* (cf. Fig. 1.1) are manageable tasks for the human imagination, but are out of the league for currently used generative image models [3, 4], that typically rely on non-interpretable vector spaces with entangled dimensions. Building imagination machines has been proposed as a goal for artificial intelligence research recently [5]. To

imagine, is to manipulate of an internal model to generate internal images. In this sense, in the context of generative modelling, disentangling factors could as well lead the way from a science of images to a science of imagination.

1.2 How not to Disentangle.

Figure 1.2: The image captions are generated by a deep neural network (Neuraltalk2) [6]. Yet, common sense understanding of psychological and physical entities in terms of causal relationships and narratives is absent [7]. Instead, the neural network seems to capture mere associations.

Can machines tell a story? Carefully observe your own mind, when viewing the images shown in Fig. 1.2: observe how the human mind immediately interprets and jumps to conclusions, tries to tell itself a story that explains an image, whereas the machine (in this case, NeuralTalk2 [6]), is comically descriptive in contrast. The missing *common sense* may be due to a missing causal reasoning, due to a missing disentangled causal representation of the world. But how to learn a disentangled representation from scratch, *i.e*. from raw image data? - As we will find out (in Sec. 2.4), disentangling causal factors from raw image data without any side information is impossible theoretically, but can only work with model assumptions or interventional data.

Let us consider an example to visualize the fundamental problem: Given an image dataset of human persons, that has strong variation in the pose and in the appearance of the persons, how to find these two underlying axes of variation (pose and appearance)? Let us suppose the distribution of variation follows a two-dimensional Gaussian distribution, one dimension for pose, one for appearance. By this we assume that images are only generated from the causal factor of pose and the factor of appearance. The learning algorithm has access to randomly sampled images from this distribution. An optimal data compression algorithm will be able to fit a function from the images to the two-dimensional subspace, which explains (by assumption in this example) the variation in the dataset. But are the two dimensions, that the algorithms finds, disentangled? No. In fact, any linear combination of pose and appearance axes and its orthogonal complement are equally valid to span the subspace of underlying variation. Just from observing a two-dimensional Gaussian, no meaning will be attached to the axes. In practice, this

problem is often circumvented by first fitting a generative model to the image dataset and *afterwards* interpolating in the latent space to determine (by human judgment) the axes of interest (here the pose or appearance axis). The meaning of pose and appearance as independent factors comes from the fact, that it is easily possible in the real world to change one factor without the other. A person that walks (hence changes shape) without loosing clothes on the way (not changing appearance) is a trivial example for that.

In summary, on the basis of dataset statistics alone one cannot disentangle causal factors if the information about how to select the axes, *i.e.* which factors to separate, is not contained in the raw data. Fitting a model to the data distribution, does in general not give insight into how the data was generated.

1.3 How can Humans Disentangle?

The dichotomy between humans and machines is constructed, of course, because on a fundamental level humans are machines. But in this context, the distinction between humans and machines shall refer to the current gap between human and machine learning performance, in terms of inferring generative factors and reasoning (again, cf. Fig. 1.2). So, what advantageous characteristics does the human mind have, that are lacking in data-driven machine learning algorithms?

Priors. Whether acquired or inherited, certain inductive priors seem to guide the human learning in its early phases [7]. Archetypal knowledge of psychology [8], a universal grammar for language [9] and causal intuitions on everyday physics [10] are some of the cognitive priors, that could explain the intuitive psychology, the rapid language acquisition and the remarkable causal inference from limited amount of data.

Data. Not only quantity, but also quality of data. Machine learning on images is commonly posed as the task to learn from randomly sampled images from a data set. But humans do not perceive the world in arbitrary samples. To humans, the world appears in a temporal sequence, which reveals how generative factors change and persevere across time. Instead of focusing on datasets with static images, sampled at random so that the images may have nothing to do with each other, algorithms should use video datasets and harness the rich temporal information. Another key difference is, that humans interact with their environment. That means, humans know how factors change, not only by observing them changing, but also by changing them. Anyone, who has watched a human infant play, can affirm that the learning mind is obsessed with interaction and change. The inevitable destruction around a young human is no accident, but a result of curious learning. Whether by performing consciously controlled experiments or by subconscious cues [11]: Interaction is crucial for a learning mind.

Models. Humans are able to imagine. That presupposes an internal model of the world, to which specific changes of representational factors can be applied. In machine learning, fitting neural network models as functions to approximate datasets has seen tremendous progress recently - to the point that it is considered a solved problem. This progress is mainly due to the effectiveness of neural networks to fit high-dimensional functions. But a probabilistic fit to a dataset, however complex and rich, is not a causal model. Even if

one were to obtain a probabilistic model over all images of the world (one could start with *e.g*. ImageNet [12]), this would tell very little about the real-world (causal) relationships between objects.

What can we learn from these differences? An algorithm to understand the world: should contain useful *prior* assumptions to efficiently use temporal and interactional *data* that contains the necessary causal relationships and interactions, to learn a useful *model* of the world.

1.4 Problem Formulation

In this thesis, we focus our efforts on image datasets which feature a single object class. Typically, objects appear in an intricate interaction of many causal factors, that can vary. For example, in an image dataset of people, the persons can have a different clothing, skin color, body physique and posture. The image generation process itself may further add factors like illumination, viewpoint or contrast. The manifold factors of possible variation in objects can be classed into to prominent categories: the objects geometric versus its visual properties. The object's geometry subsumes factors like pose, viewpoint and physique to all of which we collectively refer to as *shape*. The object's visual properties are the complement of shape. They encompass qualities like color, texture and shading and are referred to as *appearance*. Disentangling shape from appearance is a difficult problem, due to their intricate interplay, especially under heavy object articulation. Consider a person raising an arm: the color and texture of the pullover sleeve intrinsically does not change, but appears at a different location in the image. The complexity for modelling this interaction enters, as a variation in shape is a change of the domain of appearance rather than a change of its values [13]. Even simple shape models [14] offer difficult optimization problems. An efficient model for shape should cover all possible states of the object and preserve the local linkage to its intrinsic appearance. For this we partition both shape and appearance.

1.5 Contributions

This thesis provides evidence for two hypotheses:

- *Hypothesis* i)*: Unsupervised learning of object shape benefits from abstracting away the complement of shape, namely the object appearance. Explaining away the appearance factor can be achieved by a disentangled generative modelling of both factors.*

- *Hypothesis* ii)*: Learning unsupervised disentanglement without any assumptions is fundamentally impossible. In accordance with the literature on causal learning [2], disentangling causal factors requires model assumptions and/or interventional data - instead of observational (raw) data.*

To address these hypotheses, we *explain*, *validate* and *evaluate* a method for unsupervised shape learning: *Unsupervised Part-wise Disentanglement of Shape and Appearance* developed by Lorenz *et al*. 2018.

To *explain*, after theoretical prerequisites (Chapter 2) we give an overview over state-of-the-art unsupervised disentangling literature and situate the proposed method in relation to the literature (Chapter 4). In particular, we carve out the necessary aspects of an approach for disentangling causal factors and analyze the current state of research in order to indicate future directions. We subsequently disclose our method (Chapter 3).

To *validate*, we show that the proposed method outperforms the state-of-the-art for unsupervised learning of object shape on miscellaneous datasets, featuring human and animal faces and bodies (Chapter 5). We also contribute several self-made video datasets for disentangling human pose from appearance, for articulated animal motion and for articulated composite objects. We highlight the specific challenges of these datasets and elucidate how the proposed method tackles them.

To *evaluate*, we perform ablation studies on critical components of the method. In addition, we compare to a part-wise shape learning method which does make the goal of disentangling explicit. To show that the disentanglement is indeed achieved, we evaluate the disentanglement performance against a shape-supervised state-of-the-art disentanglement method and perform favorably (Chapter 6).

In short, our results are a big improvement upon the state-of-the-art in unsupervised object shape learning. This confirms the first hypothesis. To complement the learned shape in a generative process, object appearance is disentangled from shape. The achieved disentanglement with our causal assumptions, and the not-achieved disentanglement when dropping these assumptions, confirms the second hypothesis.

2 Prerequisites on Learning Disentanglement

2.1 Learning from Data

Learning from data is commonly understood as the ability of algorithms to improve their performance on a task with experience accumulated from the observation of data [15]. The source of data is usually a dataset - set of data points $X = \{x_i | i \in \{1 \dots n\}\}$, which are sampled from a probability distribution $x_i \sim p(x)$. In general, these data points are multi-dimensional. In computer vision in particular, data are images $\mathbf{x}$ with height h and width w, so that the data points are $\mathbf{x} \in \mathbb{R}^{h \times w}$.

2.1.1 Supervised

The term supervised learning denotes the task to learn a mapping from data points x_i to target labels y_i. A supervised algorithm has access to data-label pairs $(y_i, x_i) \sim p(y, x)$, in order to estimate the connection between data points and labels, either in form of a conditional probability $p(y|x)$, or in form of a deterministic function $y = f(x)$. The label y can be either discrete (*e.g.* information about an object class) or continuous (*e.g.* the location of an object part in an image). Recent advances, in particular the effectiveness of neural network models (cf. Sec. 2.1.3) on big datasets, have led to huge progress on problems that can be formulated as regression or classification. That is why on many traditional computer vision problems, such as object recognition, image classification or human pose estimation, machines are now performing on a superhuman level; hence, these problems are now considered to be essentially solved.
The Achilles' heel of supervised learning lies in the need for a viable supervision signal. To get labels, it is usually required to manually annotate the data. The human effort in this is costly, error-prone and not scalable to the ever-growing vast amounts of raw data.

2.1.2 Unsupervised

Unsupervised learning is the endeavour to learn about structures and patterns in unlabelled data. In this paradigm, the learning algorithm has access to the samples of the data distribution $x \sim p(x)$. The task is usually framed as a form of density estimation, *i.e.* to model the entire distribution in a probabilistic generative model (cf. Sec. 2.2). Unsupervised learning is considered much harder than supervised learning [16]. There are several complications in the design of unsupervised algorithms:

- Naturally, without supervision, *the goal of learning is not specified*, hence surrogate objectives have to be formulated. The lack of specification renders the evaluation oftentimes arbitrary and subjective [17].
- It is a priori not clear, *how much prior knowledge* should be embedded. To introduce no artificial bias, some argue for a purely data-driven approach. Others argue for the importance of certain inductive priors to guide learning [7]. A related modeling choice is, whether the algorithm should be model-free or model-based. In this work we argue for using more prior knowledge and modelling assumptions to obtain strong constraints.
- Lastly, the *definition of the term unsupervised* itself is subject to discussion. What entitles an algorithm to be called unsupervised? While the definition itself has no practical importance, unclear and imprecise terminology unnecessarily confuses. Here, unsupervised learning shall mean to use no label information for the dataset samples, but assuming a model or inductive priors shall be fine. This is indeed necessary to enable unsupervised learning of disentanglement at all, as we will see in Sec. 2.4.

2.1.3 Artificial Neural Networks

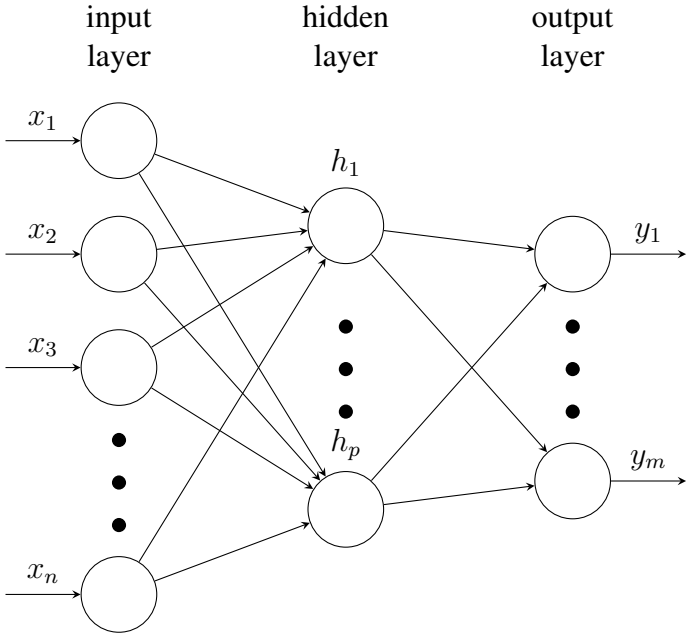


Figure 2.1: Sketch of a one-hidden-layer artificial neural network model: input $x = \{x_i | i = 1 \dots n\}$ and output $y = \{y_j | j = 1 \dots m\}$ are connected through a hidden layer $h = \{h_k | j = 1 \dots p\}$

Artificial neural networks are a powerful and flexible tool for function approximation. Inspired by biological neurons, there have been numerous questionable claims w.r.t. their

biological plausibility. Here, we will treat an artificial network solely as a parametric non-linear function approximator. They can approximate a function $y = f(x)$ with vector input $x = \{x_i | i = 1 \dots n\}$ and vector output $y = \{y_j | j = 1 \dots m\}$, also see Fig. 2.1. At minimum they connect the input and the output through one hidden layer $h = \{h_k | j = 1 \dots p\}$ by:

$$\begin{aligned} h_j &= a(\sum_i w_{ji} x_i + w_{0i}) \\ y_j &= a'(\sum_i w'_{ji} h_i + w'_{0i}), \end{aligned} \tag{2.1}$$

with weight matrices w, w', non-linear so-called activation functions a, a' and bias vectors w_0, w'_0. Neural networks can also comprise multiple hidden layers connect by $h_j = a(\sum_i w_{ji} h_i + w_{0i})$. It can be shown, that in the limit of infinite hidden units h_j a one-hidden-layer network is enough to approximate any (continuous) function arbitrarily close [18, 19]. In practice, however, networks with more than one hidden layer, referred to as deep neural networks, seem to work better. This may be due to the possibility of building a hierarchical feature representation [20], that reflects the hierarchical nature of the physical reality. Typical activation functions are for example the sigmoid function (a_{sigm} or the rectified linear unit (a_{relu}):

$$a_{\text{sigm}}(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

$$a_{\text{relu}}(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases} \tag{2.3}$$

The activation function needs to be non-linear, otherwise the neural network is just a linear classifier (matrix multiplies are again matrices). For processing image data, the weight matrices can be constrained to be only locally connected and to share weights across locations to enforce translation invariance, resulting in *convolutional* neural networks.

Deep neural networks have highly non-convex likelihood functions, hence for optimization iterative numerical methods are used: The weights w are initialized to some initial value w^0 and then updated at time step t with an update rule $w^{t+1} \to w^t$. A simple yet successful rule is given by gradient descent,

$$w^{t+1} = w^t + \lambda \nabla_{w^t} \mathcal{L}(w^t), \tag{2.4}$$

where λ is the learning rate, parametrizing the step size. In practice, calculating derivatives of the likelihood w.r.t. the weights can be done efficiently via error backpropagation. For big datasets it becomes cumbersome to calculate the gradient w.r.t. the whole dataset. Taking only a random subset of the data for an approximation of the gradient, renders the optimization stochastic; the procedure is then called stochastic gradient descent.

2.2 Generative Models

> What I cannot create, I do not understand. - R. Feynman

Learning and understanding structure in data by being able to generate, is the rationale behind generative modelling. Generative models are mostly applied for unsupervised learning and can be contrasted to discriminative models. While discriminative models are used to model posterior conditionals $p(y|x)$ (*e.g.* for supervised learning (cf. Sec. 2.1.1), generative models capture the complete data distribution $p(x)$ in an estimate $\hat{p}(x)$ [16]. Thus, after estimation, one can generate samples from this model $\hat{p}$. Hence the name generative model. Generative modeling can for example be used for outlier search, where regions with low probability under the model are taken as indicative for an outlier. The currently predominant generative models are built on either autoencoding or adversarial formulations:

2.2.1 Autoencoding Formulations

An autoencoding model is learning by reconstructing samples of data, $\hat{x} = f(x)$. To enforce data compression (otherwise the identity function is a trivial solution of autoencoding) the function has an information bottleneck, namely an inferred latent code z of reduced dimension. The autoencoder is then the chain of an encoding function $z = e(x)$ and a decoding function $\hat{x} = d(z) = d(e(x))$.

Whereas the conventional autoencoder consists of deterministic mappings e, d, the variational autoencoder [4] models the probability distribution $p(x)$. More specifically, it maximizes a lower bound to the logarithmic likelihood $\log p(x)$ of data x. This so-called variational lower bound $\mathcal{L}$ is given by:

$$\mathcal{L} = \mathbb{E}_{z\sim q(z|x)} \log p(x|z) - \mathbb{E}_{z\sim q(z|x)} \log \frac{q(z|x)}{p(z)} \tag{2.5}$$

Where z introduces latent variables, with a prior distribution $p(z)$, with an approximation to the posterior $q(z|x)$ of the latent variables, and the posterior of the data given the latent variables $p(x|z)$. If one wants to model the distributions with neural networks, one typically uses Gaussian distributions and lets the networks predict the parameters (mean μ and variance Σ) based on the image. In the current machine learning contexts, all functions (e, d) and moments (μ, Σ) are modelled with neural networks.

2.2.2 Adversarial Formulations

Generative adversarial networks (GAN) [3] consist of two neural networks competing in a zero-sum game. A generator network G is generating images based on a latent code z sampled from a distribution $p(z)$. The discriminator network D is a binary classifier with the task to classify an image as originating from the data distribution p_{data} or from the

distribution produced by G. The loss function of G is the negative of the loss of D, such that one can formulate the optimization in a minmax form:

$$\min_D \max_G -\frac{1}{2}\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] - \frac{1}{2}\mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))] \tag{2.6}$$

The generator is then optimized to make the output indiscriminable from the data distribution. The discriminator can be interpreted as a learned similarity metric, to measure the closeness of an image to the data distribution [21]. There are many variants and extensions to this basic principle of learning with an adversarial task. For example, one can learn a discriminator for a set of image patches [22].

2.3 Disentangling Representations

In supervised learning, a performance measure is naturally induced by the metric, that is being optimized. In the unsupervised setting, judging the performance of a model is less straightforward. How to rate the quality of the latent representation?

2.3.1 Learning Representations

> Disentangle as many factors as possible, discarding as little information about the data as is practical. - Bengio *et al*. [1]

According to Bengio *et al*. [1], a representation is useful, if it can be applied to many - in advance unknown - different tasks, while being trained on only one particular task. As the downstream tasks can be multifarious, the essential *information* should be contained in the representation. For some tasks only a subset of aspects of the data will be necessary, that is why *disentangled factors* make a representation particularly practical.

The latent representation z learned by generative models captures the essential *information* of the data distribution. That is made sure by requiring the ability to generate samples from the original data distribution from it. How then to reach the second goal, the *disentanglement* of generative factors?

2.3.2 Disentangling defined by Equivariance and Invariance

What is a factor? As outlined in the introduction (cf. Sec. 1), factors in a representation should correspond to causal elements of the world. In general, these factors can interact in complicated ways to finally result in an image. Here, we only consider the case where multiple independent factors each have an influence (cf. Fig. 2.2):

$$p(z_1 \dots z_N) = \prod_i p(z_i) \tag{2.7}$$

A change in an element, should then lead to: *i)* a corresponding change in the representational factor and *ii)* leave other factors, that represent other elements, unchanged.

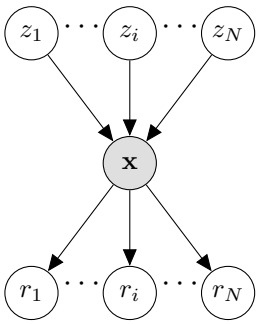


Figure 2.2: Disentangling causal factors means to infer an estimate - *i.e*. a representation - from an image

Formally, this can be seen as inference: a number of latent variables $z_1 \dots z_N$ interacted to cause the existence of the observed image $\mathbf{x}$. The task is now to infer estimates for these latent variables $r(\mathbf{x})_i := r_i$. A graphical model of the process is shown in Fig. 2.2. A disentangled representation should simultaneously fulfill equivariance and invariance: A change in z_i should: *i) equivariantly* change in the abstract representational factor r_i, *ii)* while leaving the other factors $r_j, j \neq i$, that represent other causes, *invariant*.

2.4 Theoretical Impediments from Causality

Generative factors represent causal elements. Learning a disentangled representation of generative factors is then understood as causal inference. In accordance with the causal literature [2], we can make statements about the type of knowledge, that can be gained by the type of data provided. It turns out that from "raw" image data, it is actually impossible to learn a disentangled representation z - raw data referring to images x sampled from $p(x)$, without further assumptions. To elucidate this fact, we start with a primer for causal learning (Sec. 2.4.1), outline which inductive biases are needed for disentanglement (Sec. 2.4.2).

2.4.1 Causal Learning

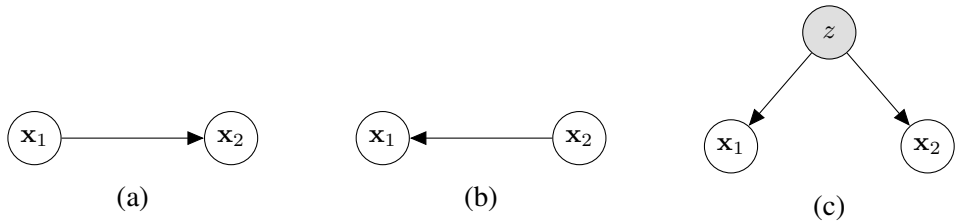


Figure 2.3: Correlation implies causation - if x_1 and x_2 correlate, a) x_1 may cause x_2, b) x_1 may be caused by x_2 or c) both are contingent on a latent cause z

Learning to infer causality is harder than statistical learning. We outline the basic problem for the case of two variables x_1, x_2: statistical learning aims at estimating probabilistic properties such as $p(x_1, x_2)$ or $p(x_2|x_1)$ from data. It is a well-known theme in statistics is that correlation does not imply causation. Less well-known is Reichenbachs principle [23, 24], that states: if two random variables are statistically dependent, then there exists a third variable that influences both or a direct causal link between them (Fig. 2.3). In addition to estimating the probability distribution, also the causal structure has to be inferred [23].

To show the limitations of raw data, we sketch an intuitive example problem (adapted from [25]): How to learn the causal connection between a barometer and the weather? If the barometer is working well, there exists a clear correlation between the weather condition and the needle position. Given a dataset showing both barometer and corresponding weather condition, a capable machine learning algorithm will be able to capture this correlation. However, it will fail to understand the causal direction, since this is not possible from the data. Imagine how a human would go about solving this problem: Having a mechanistic model of the world he could reason about the precise causal mechanism relating weather to air pressure to needle position. A simple model could be: weather influences air pressure, pressure influences barometer needle position. What if one has no prior knowledge? A solution of child-level simplicity is, to force the needle to move with a finger. Without the power of magic, the weather will not change. Hence causality has to go other way or via a third latent variable influencing both *i.e.* air pressure. To conclude, the strength of association (correlation) can be estimated with observational data alone, this can answer the question: how likely will it rain, if the barometer needle sinks? But not: how would the weather change if I force the barometer needle to sink?

Pearl [25] distinguishes between three types of questions, that can be answered by different types of knowledge:

Table 2.1: Ladder of causation [2]. Questions at level i of the ladder are only accessible with information from level i or higher.

Level	Symbol	Typical Activity	Typical Questions
1. Association	$P(y\|x)$	Seeing	What if I see?
2. Intervention	$P(y\|\mathrm{do}(x), z)$	Doing	What if I do?
3. Counterfactual	$P(y_x\|x', y')$	Imagining	What if had done?

The levels of this *ladder of causation* [25, 2] are separate not only conceptually, but in the type of data or assumptions that have to be made in order to access them. In particular, by unsupervised learning from observational data only the first level is accessible. The second level requires interactional data or model assumptions, while the third is inaccessible without an explicit model. The answers to these hypothetical questions (counterfactuals) lie by definition not in the data (facts).

2.4.2 Disentangling requires Interventions or Model Assumptions

The results from the study of causal inference also entail that "purely" unsupervised disentangling, *i.e.* estimating $\hat{\mathbf{z}}_\mathbf{i}$ from samples $x \sim p(x)$, is impossible. A proof for this can be found in [26]. Current machine learning operates mostly on the level of association, estimating (complex) correlations from raw data. As we have seen, this purely data-driven approach can only go so far. In contrast, humans seem to have the ability to interact with their environment and have innate assumptions on coherence, causality, physics etc., which introduce inductive priors [7]. To bring *i)* interventions and *ii)* model assumptions to our problem of disentangling shape and appearance, we *i)* apply changes to an image, which are assumed to change only one factor and *ii)* model the causal process of the image generation in the theme of analysis-by-synthesis.

3 Method

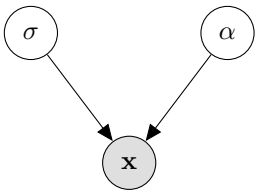


Figure 3.1: An image $\mathbf{x}$ is assumed to be generated from the factors of shape σ and appearance α. Implementing an intervention with a transformation of factors, means changeing one factor without changing the other.

To disentangle shape σ and appearance α, we can emulate interventions by image transformations. Under the assumption that certain transformations only lead to a change in shape, while leaving appearance invariant or vice versa, we obtain access to interventional data.

Additional assumptions can be made about the image generation process in the regime of analysis-by-synthesis: The key idea is, that the process of how an image is generated from underlying factors (graphics), is much better known than estimating the factors from the image (inverse graphics). One can combine a model for analysis and a model for synthesis to reconstruct an image (autoencoding, cf. Sec. 2.2.1). Herein the synthesis can be tightly constrained to fit the assumptions about reality [27]. Assumptions about how shape and appearance interact enable disentanglement. The idea to generatively entangle in order to disentangle has been explored before in other contexts [28].

To capture an object in an abstract representation, we follow two key ideas: *(i)* disassembling the object into its constituent parts and *(ii)* disentangling spatial geometry (shape) from visual features (appearance). Hence, we model an object as a composition of parts, each part with a part appearance and a part shape, as sketched in Fig. 3.2. The part shape should correspond to the area in the image where the part is located, whereas the part appearance is a feature descriptor for that area. The overall object representation is then the collection of part shapes and part appearances.
The disentanglement of shape and appearance can be enforced by demanding that shape is invariant under the transformation of appearance and vice versa. This is realized in a two-stream autoencoding formulation. Here, an image is reconstructed from a combination of shape and appearance, with shape extracted from the appearance-transformed image and appearance from a shape-transformed image. Additionally, the part shape is tied to the location of the part in the image: an equivariance loss encourages that the part shape moves in unison with the part in the image. We implement these objectives into a loss framework, which is explained in sec. 3.1.

To assert a decomposition into independent local parts, we ensure their local modelling and treatment throughout the whole pipeline. This is highlighted when describing the architecture in sec. 3.2.

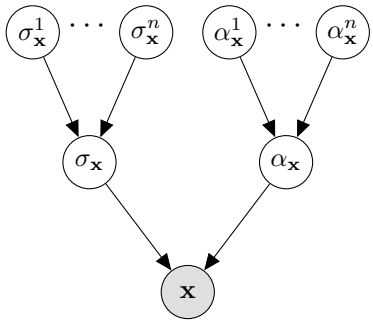


Figure 3.2: Modelling an image $\mathbf{x}$ of an object with shape $\sigma_{\mathbf{x}}$ and appearance $\alpha_{\mathbf{x}}$, by factorizing into part shapes $\sigma^i_{\mathbf{x}}$ and part appearances $\alpha^i_{\mathbf{x}}$

3.1 Disentangling by Transforming

We want to represent an object in an image $\mathbf{x}$. Let us denote the part shape for part i with $\sigma^i_{\mathbf{x}}$ and the part appearance with $\alpha^i_{\mathbf{x}}$. For an object with n parts, the overall shape is constituted by the collection of its part shapes $\sigma_{\mathbf{x}} = (\sigma^1_{\mathbf{x}}, ..., \sigma^n_{\mathbf{x}})$, the same goes for the appearance $\alpha_{\mathbf{x}} = (\alpha^1_{\mathbf{x}}, ..., \alpha^n_{\mathbf{x}})$. We model the part appearances as feature vectors, the part shapes are chosen to be scalar fields like the image itself. Thereby one can establish a direct correspondence of locations in the image to locations in the shape representation.
How do we disentangle the shape and appearance components in the representation? In general, a variation in shape will not affect appearance and vice versa. Thus, if we deliberately change shape without changing appearance, we can enforce the invariance of the appearance representation under such a change. We refer to these changes as shape transformations $s : \mathbf{x} \to s(\mathbf{x})$, which, if applied to an image $\mathbf{x}$, directly act on the underlying pixel space Λ. Along the same lines we can define appearance transformations $a : \mathbf{x} \to a(\mathbf{x})$, which act on the image itself. The shape should be invariant under change of appearance, conversely, the appearance should be invariant under change of shape. In addition, the shape should transform in the same manner as the image. That means the shape representation is assumed to be equivariant under shape transformations. In summary:

$$\alpha_{s(\mathbf{x})} = \alpha_{\mathbf{x}} \qquad \text{(invariance of appearance)}$$
$$\sigma_{a(\mathbf{x})} = \sigma_{\mathbf{x}} \qquad \text{(invariance of shape)}$$
$$\sigma_{s(\mathbf{x})} = s(\sigma_{\mathbf{x}}) \qquad \text{(equivariance of shape)}$$

Our method builds on the autoencoding paradigm, with part shapes and part appearances assuming the role of the latent code. To incorporate these constraints into the loss of an autoencoder, we reconstruct an image $\mathbf{x}$ not from the shape and appearance

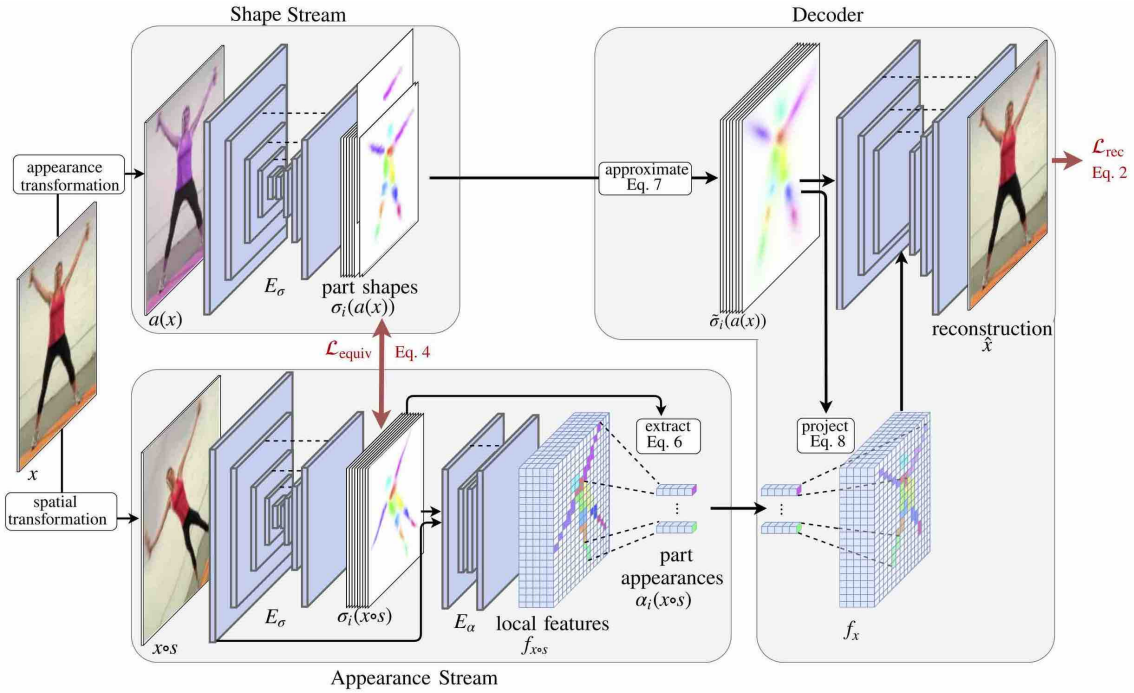


Figure 3.3: Encoder E encodes shape and appearance for two transformed images $s(\mathbf{x})$ and $a(\mathbf{x})$, after recombination R of $(\alpha_{s(\mathbf{x})}, \sigma_{a(\mathbf{x})})$ into latent image Z, the decoder D reconstructs the image $\mathbf{x}$.

$(\alpha_{\mathbf{x}}, \sigma_{\mathbf{x}})$ determined from the original image $\mathbf{x}$, but from appropriately transformed images $(\alpha_{s(\mathbf{x})}, \sigma_{a(\mathbf{x})})$. If the invariance constraints, as formulated above, are full-filled, these transformations do not change the latent code. Thus, the loss implicitly enforces invariance. To obtain shape and appearance, we encode both $a(\mathbf{x})$ and $s(\mathbf{x})$ with an encoder E. And, after a recombination (for details see sec. 3.2) to a latent image Z, a decoder D reconstructs the image. This configuration is depicted in Fig. 3.3, the reconstruction loss $\mathcal{L}_{\text{rec}}$ is as follows:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \mathrm{D}[\alpha_{s(\mathbf{x})}, \sigma_{a(\mathbf{x})}]\| \tag{3.1}$$

Let us examine what this formulation means on the level of a single part: the part appearance $\alpha^i_{\mathbf{x}}$ is extracted at locations in the spatially transformed image $\sigma^i_{s(\mathbf{x})}$, but then used for reconstruction at the location in the original image $\sigma^i_{\mathbf{x}}$. For example in Fig. 3.3 the appearance of the arm will be extracted in a raised position, but then these features are used for reconstructing an arm in a lowered position. For this to succeed, firstly, the appearance features need to be sufficiently abstract. Secondly, part locations of the two images have to refer to the same part and track the location of it consistently. This part assignment consistency is an implicit way to improve equivariance under the shape transformations. For a known shape transformation the equivariance of shape can also be encouraged explicitly with a loss. This has been used before in the context of unsupervised landmark learning by [29, 30] as a point-wise loss on a part probability map, encouraging the exact

location of a part to transform accordingly. In our case, the part shapes shall not encode probability, but instead the spatial extend of a part. In approximation, we want the first two moments (μ, Σ) to transform correctly. Thereby the extend and orientation of the parts is penalized in addition to its mere position.

$$\mathcal{L}^i_{\text{equiv}} = \mathcal{L}^i_\mu + \mathcal{L}^i_\sigma \tag{3.2}$$

The overall loss objective is the sum of the reconstruction loss and the equivariance loss for all n parts:

$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}^i_{\text{equiv}} + \mathcal{L}_{\text{rec}} \tag{3.3}$$

3.2 Analytic Disentangling by Synthetic Entangling

The autoencoding pipeline consists of analysis into factors and subsequent synthesis. The **analysis** is the encoding of both shape and appearance for each part. The **synthesis** is the meaningful recombination of this information into a latent image and the decoding of this latent image to reconstruct the image. Throughout the procedure we maintain the local correspondence between the representation and the image: We ensure a local appearance extraction in the encoding, a local synthesis in the recombining and a local usage of the latent image in the decoding. These architectural restrictions enable a disentangled part representation with the interpretation of a part as a localized entity.

3.2.1 Analysis

The encoding of shape and appearance given an image $(\alpha, \sigma|\mathbf{x})$[1] proceeds in two steps:
(i) $(\sigma|\mathbf{x})$: The part shapes are predicted given the image. To extract part shapes we use an hourglass neural network. We utilize the hourglass in both steps, as this model is able to preserve pixel-wise locality, and also integrates information from multiple scales [31]. The network input is an image $\mathbf{x}$, the output a stack of n part shapes $\sigma = \{\sigma^i | i = 1, ..., n\}$.
(ii) $(\sigma|\alpha, \mathbf{x})$: The part appearances $\alpha = \{\alpha^i | i = 1, ..., n\}$ are predicted given the image and the part shapes. Again we use an hourglass network, albeit shallower. The input is the original image concatenated with the stack of part shapes. The output is a feature stack F. A part appearance is obtained by averaging the feature stack with the a part shape:

$$\alpha^i = \sum_{p \in \Lambda} A(p) \frac{\sigma^i(p)}{\sum_{p' \in \Lambda} \sigma^i(p')}. \tag{3.4}$$

Each α^i now describes the appearance of a part spatially localized by the part shape σ^i.

[1] For a slim notation, we leave out the explicit reference to the generic input image $\mathbf{x}$ in this section: $\alpha, \sigma, \alpha^i, \sigma^i$ refer to $\alpha_{\mathbf{x}}, \sigma_{\mathbf{x}}, \alpha^i_{\mathbf{x}}, \sigma i_{\mathbf{x}}$.

3.2.2 Synthesis

In the analysis-by-synthesis regime, once the object representation is successfully factorized, one can make assumptions on how the factors reunite to generate an image, following the knowledge and intuition about how objects give rise to images in the physical world.

Firstly, we re-merge shape and appearance into images of descriptors at the correct locations. For each part, appearance is multiplied with the corresponding shape, yielding n part feature images:

$$z^i(\mathbf{x}) = \sigma^i(\mathbf{x}) \cdot \alpha^i. \tag{3.5}$$

Secondly, we reassemble the object from its parts: the part feature images z^i are summarized by summing in a single image:

$$Z(x) = \sum_i \frac{z^i(\mathbf{x})}{1 + \sum_j z^j(\mathbf{x})}. \tag{3.6}$$

The result is an image of part feature descriptors located according to their corresponding part shape, which we call latent image Z. Finally, the latent image needs to be decoded to an image. This is done by a neural network decoder. The decoder architecture is modeled after the up-sampling stream of a standard U-Net [32]. The latent image is scaled to different resolutions and inserted, after each layer, in addition to the part shapes. As before, the crucial property of the parts that needs to be conserved is their local direct correspondence to the image. On the one hand, one needs to assure, that the receptive field of the neurons does not extend to the full image, in order to thwart a complex non-local interaction of part information. This is why we use only half of a U-Net instead of a complete U-Net or an hourglass architecture. On the other hand, it is essential to regularize the information already before passing it to the decoder. Keeping in mind that the part shape should be of rather simple geometry, we introduce a differentiable information bottlenecks, in order to prevent the shape from being scattered over the object. It is an approximation of the part shape as

$$\hat{\sigma}^i(x) = \frac{1}{1 + (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}, \tag{3.7}$$

where μ and Σ are the mean and the covariance matrix of the part shape σ^i. This allows to pass second-order information such as size and orientation of the part to the decoder. Note that all operations are fully differentiable, such that a gradient-based optimization is possible.

3.3 Implementation Details

The image resolution is 128×128, but the resolution of corresponding part shapes is 64×64. For the reconstruction loss $\mathcal{L}_{\text{rec}}$ we use the L_1 or L_2 distance. To prevent parts

from trying to explain the whole image, instead of focusing on the object, we also restrict the reconstruction loss to an area around the part shape: a sum of Gaussian approximations around the means of the part shapes is folded with the loss.
In the decoder, the latent image Z is not only re-scaled, but also filled with parts incrementally. At the lowest scale only some parts are inserted, with each scale parts are added until at the highest scale all parts are used. This makes the part decoding a hierarchical process. The underlying assumption is, that parts exist at multiple scales. For landmark learning, we approximate the part shapes in the decoder in the bottleneck also with eq. 3.7, but fix the covariance Σ to be the identity matrix. Hence, effectively only information about the mean of each part shape can reach the decoder. This mean information is used as a landmark, so encouraging an accurate estimation of the mean through reconstruction is wanted.
To instantiate shape transformations s, one needs image pairs that show the same object in a different articulation or position: For static images an artificial thin-plate spline transform (TPS) can be applied, which generalizes rotation, scaling, translation. For video data adjacent frames exhibit natural shape transformations. The appearance transformation a is encompassing a color augmentation, contrast variations, and changes in brightness. In general, the more selective the transformation distinguishes shape and appearance, the more invariant the representation.

4 Review of Related Literature

In the following, we analyze the computer vision literature on disentangling generative factors w.r.t. the causal inference insights (cf. sec. 2.4). In this thesis the focus is on disentangling the factors of object shape and appearance and therefore we also review the research on unsupervised shape learning.

4.1 Analysis-by-Synthesis

Analysis-by-synthesis is a theme that originates from the research on language perception *e.g.* [33], but has also been successfully applied to visual perception. The idea is, to guide the cognition (analysis) by a model for generation (synthesis). This makes sense in vision particularly, as the domain-specific knowledge about the physics of image generation exceeds the knowledge about computational cognition. The theme can be realized in the form of autoencoder models [27]. Hence, the domain-specific knowledge is used to constrain the decoder, while the encoder is generic. For disentangling generative factors the analysis-by-synthesis scheme has been applied to computer vision earlier [34, 35, 28], however with the use of label information for the factors. In contrast, we apply image transformations to emulate the labels. Additionally, we choose a specific model for the interaction of shape and appearance with a local part-based model on shape and a corresponding part appearance that is linked to a part location.

4.2 Disentangled Generative Models

Closely related to the analysis-by-synthesis theme is generative modelling. Capturing essential information about data in a representation by being able to generate it is the rationale behind generative modelling. Currently the approaches in this direction are defined by adversarial [15] and autoencoding [4] model formulations. As argued in the beginning (cf. Sec. 1) and in multiple other works [28, 1, 36, 37, 38], in order to gain a conceptual understanding of the world, disentangling the underlying factors of variation is a crucial step. Recently, the endeavour for disentangling explanatory factors in the latent representation of generative models is being made explicit in the objective functions [39, 36, 37] of these models. Empirically, so far, these attempts are limited to rigid objects without articulation and disentangle holistic image factors like illumination, object rotation or total shape and global appearance. Theoretically, since these models do not make any assumptions or use interventional data they will not discover causal factors but statistical correlations (also pointed out by [40]). This suggests that from the causal perspective this line of work may ultimately prove to be futile [26].

4.3 Part-based Representation Learning

Describing an object as an assembly of parts is a classical paradigm for learning an object representation in computer vision [41] with linkage to human perceptual theories [42]. What constitutes a part, is the defining question in this scheme. Defining parts by e.g. (*i*) visual/semantic features (object detection), or by (*ii*) geometric shape, behavior under (*iii*) viewpoint changes or (*iv*) object articulation, in general leads to a different partition of the object. Recently, most part learning has been employed for object recognition, such as in [43, 44, 45, 46, 47, 48]. To solve such a discriminative task, parts will be based on the semantic connection to the object and can ignore their spatial arrangement and articulation of the object instance. Our method instead is driven by a generative process and aims at more generic modeling of the object as a whole. Hence, parts have to encode both spatial structure and visual appearance accurately. To our best knowledge unsupervised part learning and the proposed split in shape and appearance description for a part has only been used in pre-deep learning approaches [41, 49, 14].

4.4 Unsupervised Learning of Object Shape

There is an extensive literature on landmarks as compact representations of object shape. Most approaches, however, make use of manual landmark annotations as supervision signal. Humans are interested in humans, therefore most landmark annotations have been gathered for human faces [50, 51, 52, 53, 54, 55, 56] and human bodies [57, 58, 59, 60, 31, 61, 62].

To tackle the problem without supervision, Thewlis *et al*. [29] proposed enforcing equivariance of landmark locations under artificial transformations of images. The equivariance idea had been formulated in earlier work [63] and has since been extended to learn a dense object-centric coordinate frame [64]. However, enforcing only equivariance encourages consistent landmarks at easily discriminable object locations, but disregards an explanatory coverage of the object.

Zhang *et al*. [30] the coverage with more generic image modelling: the equivariance task is supplemented by a reconstruction task in an autoencoder framework, which gives visual meaning to the landmarks. However, in contrast to our work, he does not disentangle shape and appearance of the object. Furthermore, his approach relies on a separation constraint in order to avoid the collapse of landmarks. This constraint results in an artificial, rather grid-like layout of landmarks, that does not scale to complex articulations. In contrast, our method disentangles shape and appearance. Hence, for optimal reconstruction, our model has to make use of the shape information efficiently, which leads to a meaningful coverage.

Jakab *et al*. [65] proposes conditioning the generation on a landmark representation from another image. A global feature representation of one image is combined with the landmark positions of another image to reconstruct the latter. Instead of considering landmarks which only form a representation for spatial object structure, we factorize an object into local parts, each with its own shape *and* appearance description. The con-

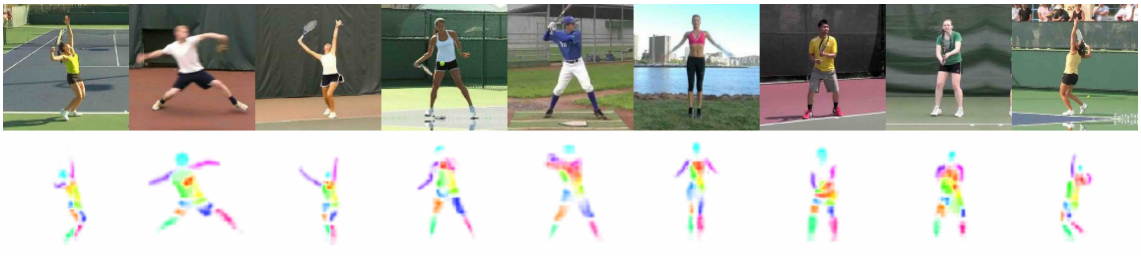
ceptual difference here is, that we understand a part not only as a point (landmark), but as an image region, with an appearance description for this region. This further factorization encourages part placement at visually meaningful locations and assists the part assignment consistency.

4.5 Disentangling Shape and Appearance

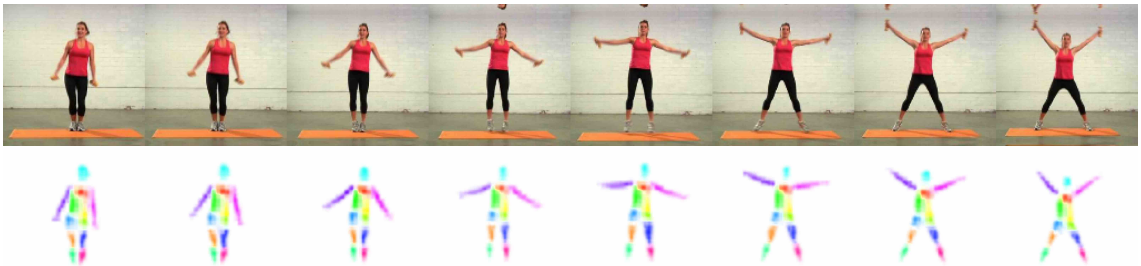
Factorizing an object representation into shape and appearance is a popular ansatz for object representation learning. Recently, a lot of progress has been made in this direction by conditioning generative models on shape information [66, 67, 68, 69, 70, 71]. While most of them explain the object holistically, only few also introduce a factorization into parts [70, 71]. In contrast to these shape-supervised approaches, we learn both shape and appearance without the explicit supervision.

For unsupervised disentangling of shape and appearance, several generative frameworks have been proposed [13, 72]. However, these works model shape itself as a surface warping and not as arrangement of parts. Therefore they use holistic models for both shape and and appearance and show results on rather rigid objects and simple datasets, while we explicitly tackle strong articulation with a part-based formulation.

5 Object Shape Learning



(a)



(b)

Figure 5.1: Learned shape representation on Penn Action. For visualization, 13 of 16 part activation maps are plotted in one image. (a) Different instances, showing intra-class consistency and (b) video sequence, showing consistency and smoothness under motion, although each frame is processed individually.

In this section we will establish that the proposed method (Chapter 3) outperforms the state-of-the-art in unsupervised object shape learning by a large margin. The learned shape representation is visualized in Fig. A.1. To quantitatively evaluate the shape estimation, we measure how well groundtruth landmarks (only during testing) are predicted from it. We obtain landmarks from our part-region based shape representation by designating the mean of a part shape $\mu[\sigma^i(\mathbf{x})]$ as the landmark position. To quantify the quality of these landmark estimates, we linearly regress them to human-annotated groundtruth landmarks and measure the test error. For this, we follow the protocol of Thewlis *et al.* [29], fixing the network weights after training the model, extracting unsupervised landmarks and training a single linear layer without bias. The performance is quantified on a test set by the mean error and the percentage of correct landmarks (PCK). We extensively evaluate our model on a diverse set of datasets, each with specific challenges.

In the following, we proceed through our shape learning results: we present the quantitative and qualitative results by object category (Sec. 5.1). On the way we introduce the

datasets for each category. In the next section we highlight and discuss the challenges, which the datasets present (Sec. 5.2) and subsequently argue for the importance of the transformations and modelling assumptions as a means to reach disentanglement and to overcome those challenges (Sec. 5.4). This confirms that disentangled modelling aids the learning of shape (Hypothesis I, Sec. 1.5).

5.1 Diverse Object Categories

We test our approach on a diverse set of object classes ranging from human and cat faces to articulated bodies and animals. In the following we go through the results sorted by object category. Where possible we compared to state-of-the-art methods quantitatively in terms of unsupervised landmark prediction, additionally we show qualitative results.

5.1.1 Human and Cat Faces

(a)

(b)

Figure 5.2: Unsupervised discovery of landmarks the object classes of (a) human (CelebA dataset) and (b) cat faces (Cat Head dataset).

For human and cat faces we use the popular datasets CelebA and Cat Head, since our predecessors for unsupervised shape learning established baselines here. Both object categories are rather rigid and non-articulated - meaning that the relations between object parts are not changing from instance to instance. Due to different breeds, the Cat Head dataset exhibits large variations between instances. Cat faces feature more complicated

Table 5.1: Error of unsupervised methods for landmark prediction on the Cat Head, MAFL (subset of CelebA) testing sets. The error is in % of inter-ocular distance.

Dataset	Cat Head		MAFL
# Landmarks	10	20	10
Thewlis *et al*. [29]	26.76	26.94	6.32
Jakab *et al*. [65]	-	-	4.69
Zhang *et al*. [30]	15.35	14.84	3.46
Ours	**9.88**	**9.30**	**3.24**

texture and locally variant silhouettes [73], hence, require a better learning of both shape and appearance.

CelebA [74] contains ca. 200k celebrity faces of 10k identities. We resize all images to 128×128 and exclude the training and test set of the MAFL subset, following [29]. As [29, 30], we train the regression (to 5 ground truth landmarks) on the MAFL training set (19k images) and test on the MAFL test set (1k images).

Cat Head [73] has nearly 9k images of cat heads. We use the train-test split of [30] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [30]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.

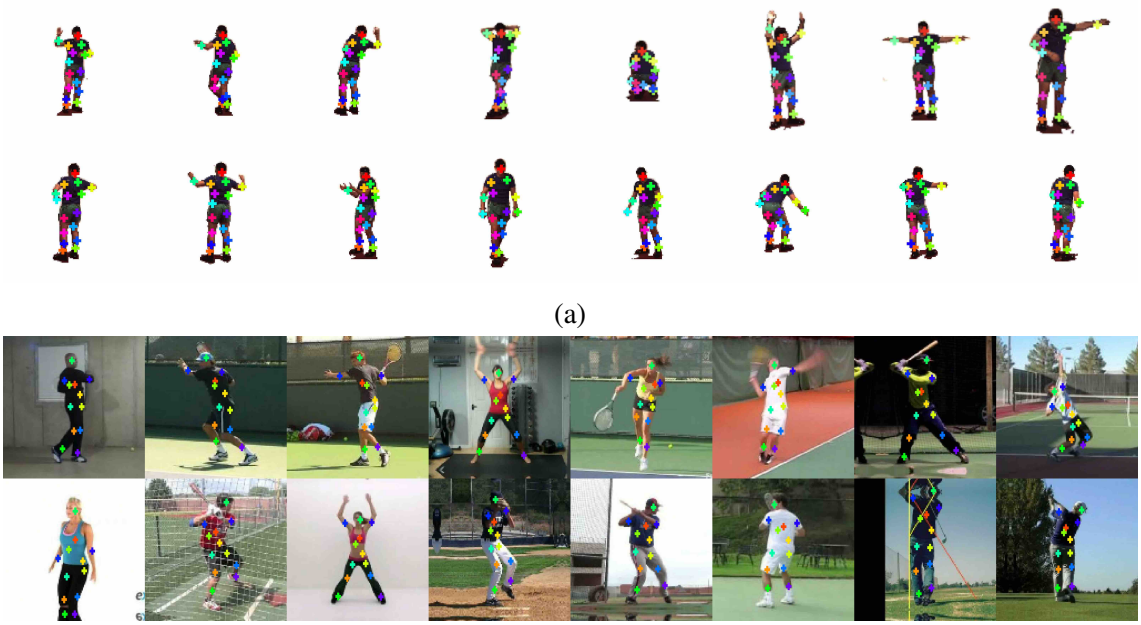
Qualitative results

The algorithm successfully defines correspondences between different human individuals, and also generalizes between different cat breeds. On both datasets the performance is visibly near-perfect. Difficulties such as out-of-plane rotation, varying lighting conditions and part occlusions (*e.g*. sunglasses) do not diminish its ability to determine the self-defined keypoints.

Quantitative results

Tab. 5.1 compares against the state-of-the-art. Our approach outperforms competing methods, with a particularly large margin of ca. $4 - 5\%$ on the more challenging Cat Head dataset. The best competitor suffers from an incomplete disentanglement (as we show in Sec. 5.3.1). An interesting side note: the most severe failure modes for Cat Head were human labelling errors, which suggests that the unsupervised performance could be better than the human labelling in this circumstances.

5.1.2 Human Bodies



(a)

(b)

Figure 5.3: Unsupervised discovery of landmarks the object classes of human bodies (a) in constrained (Human3.6M dataset) and (b) unconstrained environments (Penn Action dataset).

Human bodies introduce the challenge of object articulation. We test on three datasets. The BBC Pose dataset is used to compare against Jakab *et al*. [65], the Human3.6M dataset to beat the benchmark of Zhang *et al*. [30]. Additionally, we present the first unsupervised results on Penn Action, which is, as we argue, significantly more difficult.

BBC Pose [75] contains videos of sign-language signers with varied appearance in front of a changing background. The test set includes 1000 frames and the test set signers did not appear in the train set. Like [65] we loosely crop around the signers: we used image patches measuring 300×300 pixels, which we resized to a resolution of 128×128. For evaluation, as [65], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M [76] features human activity videos in stable environments. We consider a subset of ca. 800k images from 6 activities (direction, discussion, posing, waiting, greeting, walking). We adopt the training and evaluation procedure of [30]: We trained on 6 subjects IDs (S1, S5, S6, S7, S8, S9) and used S11 for testing. For proper comparison to [30] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset and cropped roughly around the provided bounding boxes, then resized to 128×128. For evaluation, as [30], we report the minimum of the errors to the original landmark annotations and their left-right-flipped counterparts. This is correcting for the fact, that the model does not distinguish back and frontal views (for a discussion of this problem, cf. Sec. 5.4.3).

Penn Action [77] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

Qualitative results

We demonstrate (Fig. 5.3), that our model not only exhibits strong landmark consistency under articulation, but also covers the full human body meaningfully. Even fine-grained parts such as the arms are tracked across heavy body articulations, as are present in the Human3.6M or Penn Action datasets. Also with further complications introduced with the Penn Action dataset such as viewpoint variations, blurred limbs and partial self-occlusions we are able to detect landmarks of similar quality and coverage as in the more constrained Human3.6M dataset. Additionally, complex background clutter, as in BBC Pose and Penn Action, does not hinder finding the object.

Quantitative results

The quantitative comparisons are shown in Tab. 5.2 and Tab. 5.3: other unsupervised and semi-supervised methods are outperformed by a large margin on both datasets.

On Human3.6M, judging by the performance gap, it is questionable whether the other unsupervised method from Thewlis *et al*. [29] learned to deal with articulation at all or whether they just find a mean solution.We beat the best unsupervised result by Zhang *et al*. [30] by an improvement of 2.12%. This is not only cutting the absolute error nearly in half, but also reduces the gap between unsupervised and supervised algorithms by about 77%. Zhang *et al*. [30] additionally used optical flow to stabilize their training by forcing the landmarks to cover the object, which we (and they themselves) classified as semi-supervised. Despite this advantage, our approach is able to achieve a performance gain of 1.35% even over results obtained with optical flow supervision.

On BBC Pose, we outperform Jakab *et al*. [65] by 6.1%, which translates into a reduction of the unsupervised performance gap to supervised methods by more than half. An analysis of conceptual differences to both [30] and [65] can be found in Sec. 5.3.

Table 5.2: Performance of landmark prediction on BBC Pose test set. As upper bound, we also report the performance of supervised methods. The metric is % of points within 6 pixels of groundtruth location.

BBC Pose		Accuracy
supervised	Charles *et al*. [75]	79.9%
	Pfister *et al*. [59]	88.0%
unsupervised	Jakab *et al*. [65]	68.4%
	Ours	**74.5**%

Table 5.3: Comparing against supervised, semi-supervised and unsupervised methods for landmark prediction on the Human3.6M test set. The error is in % of the edge length of the image. All methods predict 16 landmarks.

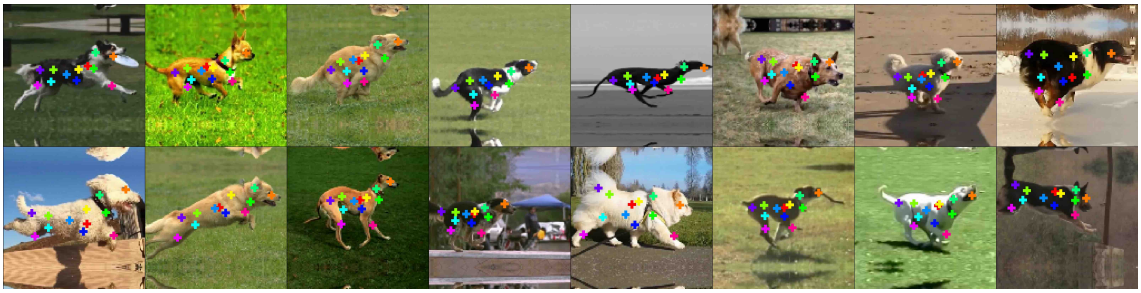
Human3.6M		Error w.r.t. image size
supervised	Newell *et al*. [31]	2.16
semi-supervised	Zhang *et al*. [30]	4.14
unsupervised	Thewlis *et al*. [29]	7.51
	Zhang *et al*. [30]	4.91
	Ours	**2.79**

5.1.3 Animal Bodies

CUB-200-2011 [78] comprises ca. 12k images of birds in the wild from 200 bird species. We excluded bird species of seabirds, because these bird species were depicted flying, (discussion in Sec. 5.2.2), roughly cropped using the provided landmarks as bounding box information and resized to 128×128. We aligned the parity with the information about the visibility of the eye landmark (discussion in Sec. 5.4.3). For comparing with [30] we used their published code.

Dogs Run is made from dog videos from YouTube totaling in 1250 images under similar conditions as in Penn Action. The dogs are running in one direction in front of varying backgrounds. The 17 different dog breeds exhibit widely varying appearances.

(a)



(b)

Figure 5.4: Unsupervised discovery of landmarks the object classes of animal bodies (a) birds (CUB-200-2011 dataset) and (b) dogs (Dogs Run dataset).

Qualitative results

On CUB-200-2011 the model discovers consistent landmarks that closely cover the birds body, despite significant variation in color, texture, size and background. The direct comparison to Zhang *et al*. [30] is further discussed in Sec. 5.3, including a qualitative comparison in Fig. 5.5. The performance on the Dogs Run dataset shows that highly differing dog breeds (size, thickness of fur) can be related via semantic parts. Furthermore, the limited amount of data in Dogs Run is no problem for finding meaningful correspondences, firstly due to the unsupervised nature of the model and the transformations acting as a form of data augmentation. On both datasets, the universality of the approach (capturing non-human poses) is underlined once more. This is to be expected, as no specific assumptions about the object-class are introduced in the model formulation.

Quantitative results

For a direct comparison to Zhang *et al*. [30] we apply their published code on the CUB-200-2011 dataset. The results are shown in Tab. 5.4. Our method surpasses [30]; to appreciate why this is so, refer to the direct comparison in Sec. 5.3.1.

Table 5.4: Error of unsupervised methods for landmark prediction on the CUB-200-2011 testing set. Both methods predict 10 landmarks.

CUB-200-2011 dataset	Error w.r.t. image edge
Zhang *et al*. [30]	5.36
Ours	**3.91**

Table 5.5: Difficulties of datasets: articulation, intra-class variance, background clutter and viewpoint variation

Dataset	Articulation	Intra-Class	Background	Viewpoint
CelebA				
Cat Head		✓		
CUB-200-2011		✓	✓	
Human3.6M	✓			✓
BBC Pose	✓		✓	
Dogs Run	✓	✓	✓	
Penn Action	✓	✓	✓	✓

5.2 Overcoming Challenges

An overview over the challenges implied by each of the presented datasets is given in Tab. 5.5. We address the main difficulties in the following: background clutter 5.2.1, intra-class variance 5.2.3, articulation and viewpoint variation 5.2.2. We make suggestions on how the method overcomes these challenges.

5.2.1 Background Clutter

The question how to separate background from object goes deeper than one might think. Fundamentally the question is equivalent to: *What is an object?* If an unsupervised algorithm is posed the task of finding the object in an image dataset - under the assumption that the object is present in all images - the object is a structure common to these images. By this definition background is everything that is not strongly correlated with the object itself. This dataset-specific object category can be unintuitive: for example for a bird sitting on a twig, the twig can considered as part of the object, if the dataset shows only birds on twigs (*e.g*. two landmarks are on feet/twig on CUB-200-2011 dataset, cf. Fig. 5.4). This dataset-biased pre-categorical thinking of unsupervised algorithms can be seen as a failure, or as a feature: on a dataset of Salsa-dancing humans our method identified the pair of dancers as an object (cf. Fig. 5.9). Technically, we allow the algorithm to focus on the reconstruction of only the object, and not background by a local weighting around the part activation (refer to Sec. 3.3). The fundamental issue of strong object-background correlations cannot be solved technically, but requires different data. Interestingly, on with the constrained but repetitive background of the Human3.6M dataset, where traditional background subtraction methods are easily applied, our method struggles: several parts

are assigned to background objects. On the other hand, complexly cluttered backgrounds - as long as no correlations to the object exist - such as the background TV screen in the BBC Pose dataset (cf. Fig. 6.3) are actually favorable for the method. This is due to the crossed reconstruction objective: if reconstructing the background of the target image is possible with the information of the source appearance image, the algorithm will try to do so by assigning parts to the background, if not, not.

5.2.2 Object Articulation and Viewpoint Variation

Object articulation and viewpoint variation makes consistent shape learning challenging. In contrast to rigid bodies that can vary in orientation and scale, articulated objects viewed from different viewpoints have many degrees of freedom more. Complex articulation can be conceptually simplified by regarding an object as a collection of rigid parts, that again each have only few degrees of freedom. This description of an arrangement of parts is exponentially cheaper than trying to capture it as a whole [41]. To enforce the factorization into independent parts it is crucial to restrain the part features and region to be local (how exactly is shown in Sec. 3). Landmarks are a simple and efficient implementation of this part factorization idea. In this sense, using landmarks as shape representation is the natural approach for tackling articulation. However, related work does not strictly enforce locality of the learned features (compare 5.3.2) or a local reconstruction from these features.

Part assignment consistency means equivariance w.r.t. changes in shape due to articulation. Equivariance is enforced twice in the method (cf. Sec. 3.1). We showed previously that the method can deal with strongly articulated human and animal bodies (cf. Sec. 5.1.2 and Sec. 5.1.3).

5.2.3 Intra-Class Variation

Intra-class variation can be both in shape and in appearance. Due to different breeds and species the animal datasets - Dogs Run, CUB-200-2011, Cat Head - present the highest degree of variability within the object class. The method in part coincidentally generalizes and in part inherently enforces in due to the data augmentation with shape and appearance transformations (see also Sec. 5.4).

5.3 Comparative Advantages

In this section we directly compare against the closest competitors on unsupervised learning of shape Zhang *et al*. [30] and Jakab *et al*. [65]. On a low-level, there are many differences to both approaches, some of which we discuss in the respective related work (5). On a high level, different is that Zhang *et al*. lack a principled way to disentangle shape and appearance (*e.g*. via transformations) (Sec. 5.3.1) and Jakab *et al*. model the appearance in a holistic encoding (Sec. 5.3.2).

5.3.1 Non-Disentangling Approach

Dataset	Human3.6M					
Actions	Directions	Discussion	Waiting	Greeting	Posing	Walking
Newell [31] (supervis.)	1.88	1.92	2.15	1.62	1.88	2.21
Zhang [30] (semi-superv.)	5.01	4.61	4.76	4.45	4.91	4.61
Thewlis [29] (unsuperv.)	7.54	8.56	7.26	6.47	7.93	5.40
Ours (unsuperv.)	**2.58**	**2.26**	**2.87**	**3.08**	**2.67**	**3.35**

Table 5.6: Comparison with unsupervised, semi-supervised and supervised methods for annotated landmark prediction on the Human 3.6M testing sets for selected actions. The error is in % regarding the edge length of the image. All methods predict 16 landmarks, from which the 32 ground truth landmarks are regressed.

Figure 5.5: Comparing discovered keypoints against [30] on CUB-200-2011. We improve on object coverage and landmark consistency. Note our flexible part placement compared to a rather rigid placement of [30] due to their part separation bias.

The method of Zhang *et al*. [30] is similar to our method, but does not disentangle shape from appearance. In Tab. 5.6 we list the detailed results of a comparison to Zhang *et al*. [30]. Fig. 5.5 provides a direct visual comparison to [30] on CUB-200-2011. It becomes evident that our predicted landmarks track the object much more closely. In contrast, [30] have learned a slightly deformable, but still rather rigid grid. This is due to their separation constraint, which forces landmarks to be mutually distant. We do not need such a problematic bias in our approach, since the localized, part-based representation and reconstruction guides the shape learning and captures the object and its articulations more closely.

5.3.2 Holistic Approach

The method of Jakab *et al*. [65] aims at disentangling shape and appearance with video information. Shape is then - similar to ours - defined by the change between consecutive frames. However, they do not model the link between local parts of shape and appearance, but use a holistic appearance embedding.

Dataset	BBC Pose				
Landmarks	Head	Wrists	Elbows	Shoulders	Avg.
Charles *et al*. [75] (supervised)	95.40	72.95	68.70	90.30	79.90
Pfister *et al*. [59] (supervised)	98.00	88.45	77.10	93.50	88.01
Jakab *et al*. [65] (unsupervised)	76.10	56.50	70.70	74.30	68.44
Ours (unsupervised)	**96.34**	**71.39**	**62.12**	**80.28**	**74.85**

Table 5.7: Comparison with supervised and unsupervised methods for annotated landmark prediction on the BBC Pose testing sets. %-age of points within 6 pixels of ground-truth is reported.

Figure 5.6: Comparison of regression results of our method (bottom rows) to [65] (top rows) on BBC Pose. For visualization by Jakab *et al*. (from their paper) ground truth is in circles and the corresponding regression in the same color. For our visualization the red dots mark the ground truth, the colored circles the regressed locations. The color coding is in terms of the error w.r.t. the image edge length.

Dataset	Cat Head
# Landmarks	20
full model	9.30
w/o $\mathcal{L}_{\text{equiv}}$	11.32
w/o $\mathcal{L}_{\text{rec}}$	35.0
w/o appearance transform	12.46
w/o shape transform	14.72

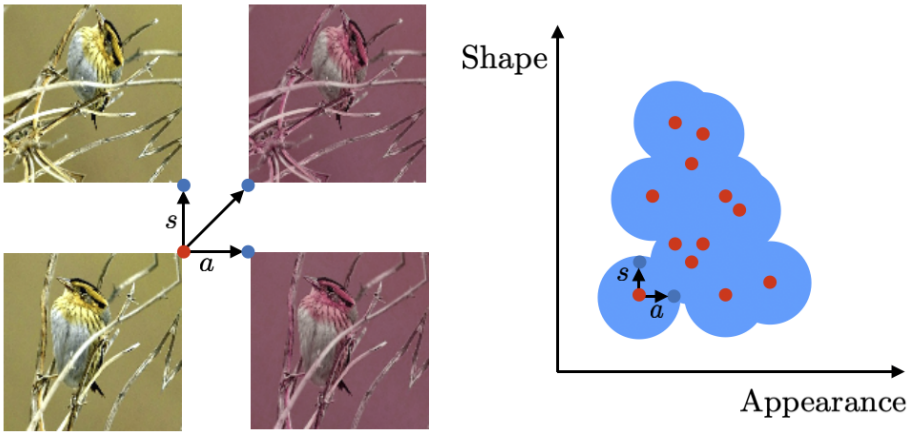
Table 5.8: Ablation studies on Cat Head dataset. We ablate the reconstruction loss $\mathcal{L}_{\text{rec}}$, equivariance loss $\mathcal{L}_{\text{equiv}}$, the color augmentation and the transformations

5.3.3 Ablating Contributions

Perhaps the most insightful comparison for a method is, to compare to lesser versions of itself: We ablate the main components of our proposed framework: reconstruction loss $\mathcal{L}_{\text{rec}}$, the equivariance loss $\mathcal{L}_{\text{equiv}}$, the appearance augmentation and the transformations for disentangling shape and appearance. For the ablation study we use the Cat Head dataset, following the already introduced train-test setup on the task of landmark ground truth regression. Tab. 5.8 illustrates the ablation results.
Leaving out the reconstruction task naturally leads to the largest drop in performance since only training on equivariance leads to collapsed landmark solutions as discussed in [30]. Note that without both shape and appearance transformations our models performs significantly worse, now comparable to Zhang *et al*. [30]. Without transformations the reconstruction objective is identical to the one of [30], only the model architecture deviates. This suggests, that the improved disentanglement (by transformations) could be explaining the overall performance gain w.r.t. [30]. Leaving out the explicit equivariance leads to the smallest drop in performance. This is not surprising, as equivariance is implicitly also enforced in the feature crossing in the framework.

5.4 Transformational Effects

In this section we discuss the effect of the transformations on learning a consistent and comprehensive representation. Since strong image transformations can make the learning curve for the algorithm too steep, we exponentially schedule the increase in magnitude, finally resulting in image changes as shown in Fig. 5.8. In effect, the transformations teach the algorithm what changes in shape and appearance are. Assuming that samples from the data distribution are - showing the same object class - related via a change in shape and appearance, the transformations blur the distribution. This data augmentation effect is sketched in Fig. 5.7.



(a) (b)

Figure 5.7: Effect of transformations on data distribution: (a) Data points (red) can be connected via a shape s and an appearance a transformation. (b) Applying transformations effectively blurs the data distribution.

5.4.1 Spatial Transformations

We perform thin-plate spline (TPS) warps to mimic spatial transformations. These changes incorporate rotation, scaling and translation as a special case. While irreplaceable for calculating the direct equivariance loss, they can result in artificial shape changes. After all, most objects - such as human beings or animals, do not warp, but articulate their parts/limbs. Natural shape changes are needed to learn a model of the objects articulation. These changes are presented in video data. Hence, for videos we enforce the reconstruction to function across different frames. This results in a much stabler performance and greater part consistency especially for highly articulated parts such as arms.

5.4.2 Appearance Transformations

Figure 5.8: Examples for shape and appearance transformation on CUB-200-2011. Images from the upper row relate to images directly below.

We mimic appearance changes with image transformations in color, contrast, hue and

brightness. Exemplars for the combined effect of spatial and appearance transformations are shown in Fig. 5.8. Especially for datasets with high intra-class appearance variance, connecting the data points via appearance changes is crucial. On Cat Head for example, without them, the method assigned different landmarks to black cats than to other-color cats. The model will incur no loss, as long as it always has to reconstruct black cats from images of black cats. If it has to relate black and white cats (*e.g.* via color inversion) this intra-class inconsistency has to vanish.
Ideally one would want more "natural" appearance changes as well. This could be a line of future work (cf. Sec. 7.1).

5.4.3 Parity Transformations

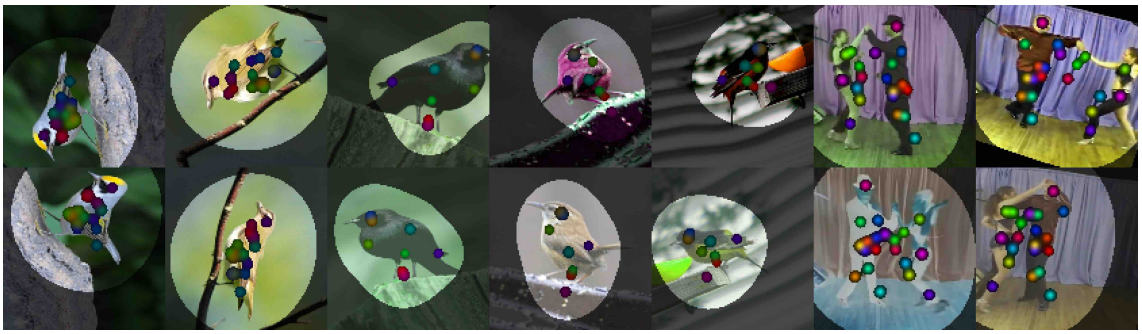


Figure 5.9: Parity changes: the images of the upper and lower row relate via the usual transformations and an additional parity flip. For the bird (1-5th column) images induced artificially, for the dancing humans (6-7th column) via sampling different frames from a video.

The model has problems with consistent part assignment under parity changes, if these changes do not change the object enough. For example human body appearance in frontal and back view are not dissimilar enough from each other. So the model can assign the same landmark to *e.g.* the right arm in the frontal view and the left arm in the back view. Similarly, there is no distinction, for dog or bird side views facing to the left or right. However, if features on the left and on the right are very different *e.g.* for the dancing humans the model automatically learns the distinction (see Fig. 5.9).

A tentative solution to the problem can be to incorporate parity flips in the equivariance loss. This is impossible for parity-symmetrical objects (such as frontal view humans), but works *e.g.* for side view dogs or birds. One has to be careful in scheduling these random parity flips, as landmarks tend to align along the mirror axis as a trivial solution. A successfully learned parity model for CUB-200-2011 is shown in Fig. 5.9.

5.4.4 Importance of Transformations

The proposed method enables to abstract away object appearance from shape. Despite the multifarious challenges in the diverse range of datasets, the method is able to learn

a dedicated part representation for shape. We compare to other approaches and reach state-of-the-art performance on the task of regressing human-annotated landmarks from the part representation. The key difference to the most competitive approach [30] is the emphasis on disentanglement via a crossed reconstruction with shape and appearance transformations. Enforcing disentanglement via targeted transformations enhances the shape representation in two ways: *(i)* it asserts that no appearance information is encoded in the shape representation and vice versa and *(ii)* it requires visual features to be equivariant under a spatial transformation. With regards to the considerations earlier, the crucial role of the transformations is to be expected, as they enable to reach a *disentanglement*. In the next section we explore and compare the performance on disentanglement.

6 Disentanglement of Shape and Appearance

Disentangled representations of object shape and appearance allow to alter both properties individually to synthesize new images. The ability to flexibly control the generator allows, for instance, to change the pose of a person or their clothing. In contrast to previous work [66, 79, 67, 69, 68, 65], we achieve this ability without requiring supervision *and* using a flexible part-based model instead of a holistic representation. This allows to explicitly control the parts of an object that are to be altered. We quantitatively compare against *supervised* state-of-the-art disentangled synthesis of human figures. Also we qualitatively evaluate our model on unsupervised synthesis of still images, video-to-video translation, and local editing for appearance transfer.

6.1 Disentangling Pose and Appearance

Figure 6.1: Transferring shape and appearance on Deep Fashion. Without annotation the model estimates shape, 2nd column. Target appearance is extracted from images in top row to synthesize images. Note that we trained without image pairs only using synthetic transformations. All images are from the test set.

We compare the disentangling performance quantitatively against a supervised method, namely the variational U-Net [66]. We evaluate on the Deep Fashion [80, 81] dataset, a

standard benchmark dataset for supervised disentangling methods. In this dataset appearance is defined by a persons ID and shape by the pose of the person. For each ID in the test set, we condition the image generation on 8 different poses which are chosen randomly from the test set. Both VU-Net and our model are conditioned on the exactly the same pose-ID image pairs. Fig. 6.1 shows qualitative results. We quantitatively compare against supervised state-of-the-art disentangling [66] by evaluating *i)* invariance of shape against variation in appearance by the distance in pose between generated and pose target image and *ii)* invariance of appearance against variation in shape by the re-identification error.

Deep Fashion [80, 81] consists of ca. 53k in-shop clothes images in high-resolution of 256×256. We selected the images which are showing a full body (all keypoints visible, measured with the pose estimator by [62]) as full visibility of the object is an assumption to the model and used the provided train-test split. For comparison with Esser *et al*. [66] we used their published code.

6.1.1 Pose Estimation

Table 6.1: Percentage of Correct Keypoints (PCK) for pose estimation on shape/appearance swapped generations. α is pixel distance divided by image diagonal. Note that [66] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [66]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

To evaluate shape, we extract keypoints using a pose estimator [62]. Tab. 6.1 reports the difference between generated and pose target in percentage of correct keypoints (PCK). As would be expected, VU-Net performs better, since it is trained with exactly the keypoints of [62]. Nevertheless, our approach achieves an impressive PCK without supervision underlining value of the embedding of object shape and the disentanglement of appearance and shape. Despite random variation in appearance the shape does not change, this can also be directly observed from the conditioned generations in Fig. 6.1.

6.1.2 Person Re-Identification

Person re-identification (ReID) is a research field on its own (overview in *e.g*. [82, 83]), the goal being to learn a similarity metric for a persons appearance, invariant to a persons posture and the image viewpoint. The key applications are automated person tracking and surveillance [84]. For our purposes, we will treat a ReID algorithm as a metric for measuring the preservation of appearance as well as the invariance to shape on our generated images. For this, we fine-tune an ImageNet-pretrained [12] Inception-Net [85] with a

ReID algorithm [86] via a triplet loss [87] to the Deep Fashion training set. On the generated images we report the standard metrics for ReID, mean average precision (mAP) and rank-1, -5, and -10 accuracy. The first question we ask, is, if the appearance encoding is stable to variations in pose, hence invariant to pose (shape). Each ID from the test set is generated in 8 different poses. The task for the ReID algorithm is now to rank the similarity of these pose-changed yet same-appearance generations. Although our approach is unsupervised, it is competitive compared to the supervised VU-Net [66] as shown in Tab. 6.2. The high chance of re-identifying a persons appearance in a different shape (rank-1 accuracy) shows that the appearance is invariant against variation in shape (pose) for both methods. To visualize the closeness of the same-ID generations in the ReID-embedding the show a t-SNE plot in Fig. 6.2.

Table 6.2: Mean average precision (mAP) and rank-n accuracy for person re-identification on synthesized images after performing shape/appearance swap. Input images from Deep Fashion test set. Note [66] is supervised w.r.t. shape.

	mAP	rank-1	rank-5	rank-10
VU-Net [66]	88.7%	87.5%	98.7%	99.5%
Ours	90.3%	89.4%	98.2%	99.2%

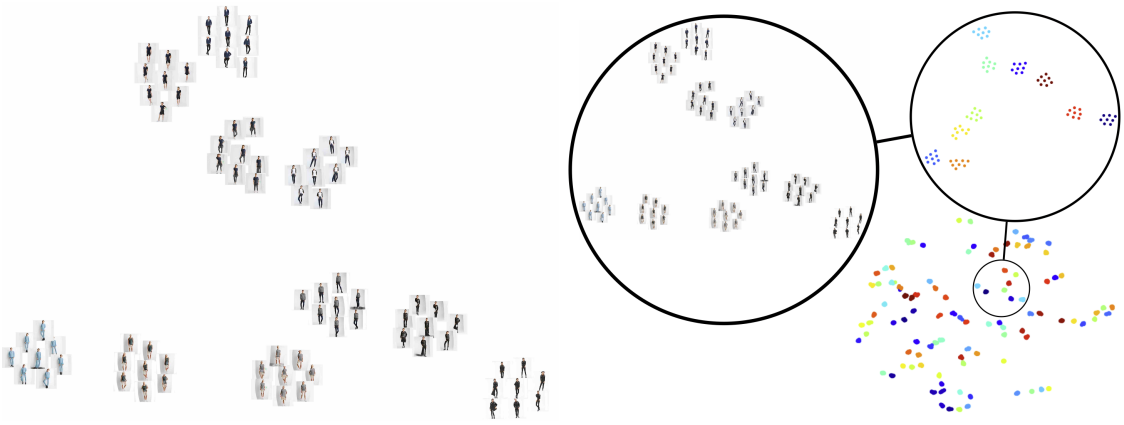


Figure 6.2: Visualization of feature distribution for generated person IDs. (Right) t-SNE (perplexity 16) of 10 generated IDs, (left) color-coded t-SNE (perplexity 12) for 10, 15, 20 and 100 IDs. Each ID has 8 samples. The different IDs are clearly separable, despite variation in pose: Hence, generated appearance is invariant to pose.

The second question one could ask is, if appearance is preserved, *i.e.* if the ReID algorithm is able to re-identify the groundtruth appearance image from the generation. Results for this are shown in Tab. 6.3. The result depends strongly on whether the algorithm had been fine-tuned to the Deep Fashion image distribution or the Deep Fashion and the synthesized image distribution. The stark difference can be explained by the difference in the feature distribution: high-frequency details (such as patterns and texture of clothing),

are not synthesized correctly, as the model is trained by a reconstruction objective which will blur these high frequencies. On the other hand, the adversarial objective will encourage *some* high frequencies, but not necessarily the ones from the initial appearance conditioning. The ReID algorithm, if not additionally adjusted to this, will pay attention to those details and subsequently fail. unambiguous function between generations and groundtruth can be found.

Table 6.3: Mean average precision (mAP) and rank-n accuracy for person re-identification from synthesized to ground truth appearance images after performing shape/appearance swap. When only fine-tuning the ReID algorithm on Deep Fashion, results are much worse that when also adjusting to the synthesized images.

Fine-tune to:	mAP	rank-1	rank-5	rank-10
Deep Fashion	17.2%	25.4%	48.8%	60.4%
Deep Fashion and Synthesized Images	75.0%	73.8%	89.5%	92.5%

6.2 Disentangling across Time

Figure 6.3: Generation results for conditioning appearances (top row) on pose (bottom, rightmost) on BBC Pose. Note that even fine-grained details in shape, such as fingers and facial expression are accurately captured.

Conditional image generation can also be extended to the task of video-to-video translation. The two conditioning images can be frames from different videos. One frame is acting as the appearance conditioning and the other as shape conditioning. By generating

each frame conditioned on the shape and appearance from two videos, one effectively transfers the appearance of one video to the shape of the other on a frame-to-frame level. We evaluate this frame-to-frame video translation on the BBC Pose dataset. The datasets videos of sign language present a delicate and complex articulation of arms and hands. We condition on appearance from videos in the training set and on shape from videos in the test set. A sample for generated frames is shown in Fig. 6.3, for the complete videos please refer to the supplementary. We want to point out two features of the model here: Firstly, despite no use of smoothing or interpolation between frames the generated sequence is smooth in the temporal domain. This is enabled by a temporally consistent part assignment which is stable across articulation. Secondly, the training on the natural spatial transforms in video data enables the model to encapsulate realistic transitions such as out-of-plane rotation and complex 3D articulation of *e.g*. hands and even fingers (note the correct translation of the thumbs position in Fig. 6.3).

6.3 Disentangling in Parts

The second type of disentanglement we approach is to factorize the object into local parts. Obviously, object parts are in general not disentangled factors in a sense that they have an independent probability distribution. Since the parts are geometrically connected, in their spatial layout they are conditioned on each other. For example, you cannot move your head arbitrarily far away from your shoulders. Still, there is a local freedom and modularity - especially in appearance features - that renders a local factorization efficient. As an illustration, the color of the shoes you wear need only be mildly correlated with your hair color. We show that the model disentangles these local modes of variation for a persons appearance (Sec. 6.3.1) and shape (Sec. 6.3.2).

6.3.1 Part Appearance Transfer

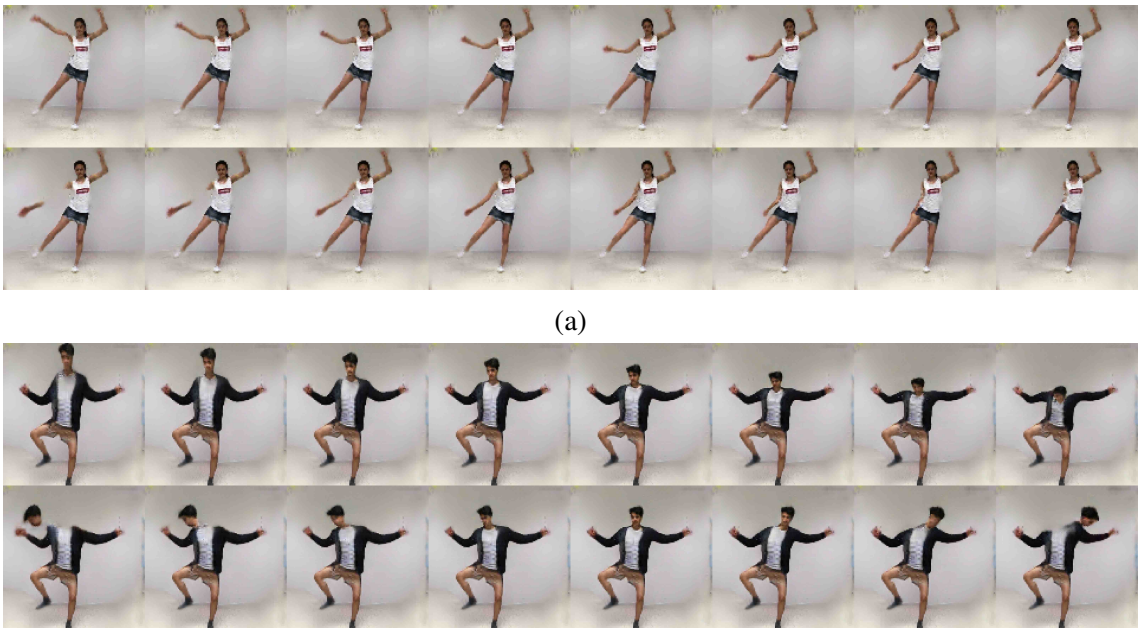
The local modelling of parts allows for a part-wise transfer of appearance. In Fig. B.1 we show the image generation conditioned on a target shape and appearance from a single image, but for several parts the appearance is transferred from another image. This shows a possible application as a virtual try-on generation, as in [88].

6.3.2 Part Shape Changes

One can also change the position of individual parts in the shape conditioning, which leads to generations as shown in Fig. 6.5. One can observe that the other non-moved parts shapes also lead to stationary parts in the generation, indicating that these parts are spatially disentangled. In the unnatural (never seen in data) regime *e.g*. if the head is to far from the shoulders, the model still hallucinates a head next to the body - similar to supervised results [68].

(a) (b) (c) (d)

Figure 6.4: Swapping part appearance on Deep Fashion. Appearances can be exchanged for parts individually and without altering shape. We show part-wise swaps for (a) head (b) torso (c) legs, (d) shoes. All images are from the test set.



(a) (b)

Figure 6.5: Moving individual body landmarks for conditional generation: (a) arm (b) head.

7 Conclusion

We have presented an unsupervised approach to learning the compositional part structure of objects by disentangling geometric shape from visual appearance. We derive invariance and equivariance constraints that enable a generative framework to discover consistent landmarks without requiring prior assumptions on landmark layout. Experiments show that our approach significantly improves upon previous unsupervised methods. Disentangling shape and appearance has been presented in the broader context of learning the causal structure of the world through images. The insights from the causal literature let us rethink the role of priors, models and data and give a direction for future work.

Throughout this work we alluded to two themes: *i)* improving models with realistic constraints and assumptions *ii)* extracting value from richer data, *i.e.* interventional and temporal data. For our task of disentangling shape and appearance of objects these themes translated into *i)* better modelling of the synthesis side in the analysis-by-synthesis framework and *ii)* utilizing image transformations w.r.t. which to capture invariant and equivariant factors.

7.1 Future Work

With regards to these two themes there are obvious improvements to be made:

i) On the modelling side our method models the interplay between shape and appearance of a composite object, but in a prototypical manner. Realistic graphical simulation - as long as it is fully differentiable - such as used in [34, 27] would impose tighter constraints onto how the factors generate the image.

ii) On the data side a next step could be to model video data in the exact temporal sequence, not only on a frame-by-frame level (cf. Sec. 6.2). To do this, the temporal changes of shape would be necessary to be modelled. For this it could prove useful to make our model generative. Generating appearance features could be implemented with standard variational features [4]. Generating shape for temporal sequences could use some type of recurrent architecture. We also repeatedly stressed the importance of the image transformations. For disentangling they are the necessary condition. The better the transformations separate variation in the to-be-disentangled-factors, the better disentangled will these factors be. Video data are the best source of shape transformations, for appearance however, the global contrast, brightness and hue transformations are neither natural nor complete of any type of appearance transform. Usually patterns and texture are considered as appearance, hence, for completeness they should also be transformed. This could be tried with a soft form of style alteration via style transfer [89]. In addition to extending appearance transformations to a higher level, one can also make them more local, this would further encourage the factorization into local parts.

Part I

Appendix

A Landmark Results

Part Activations. In Fig. A.1 we show part activation maps on video sequences from the Penn Action dataset.

Landmark Discovery. We present unsupervised landmark discovery results on the following datasets: Cat Head (Fig. A.2), Dogs Run (Fig. A.3), CUB-200-2011 (Fig. A.4), CelebA (Fig. A.5), Human3.6M (Fig. A.6) and Penn Action (Fig. A.7).

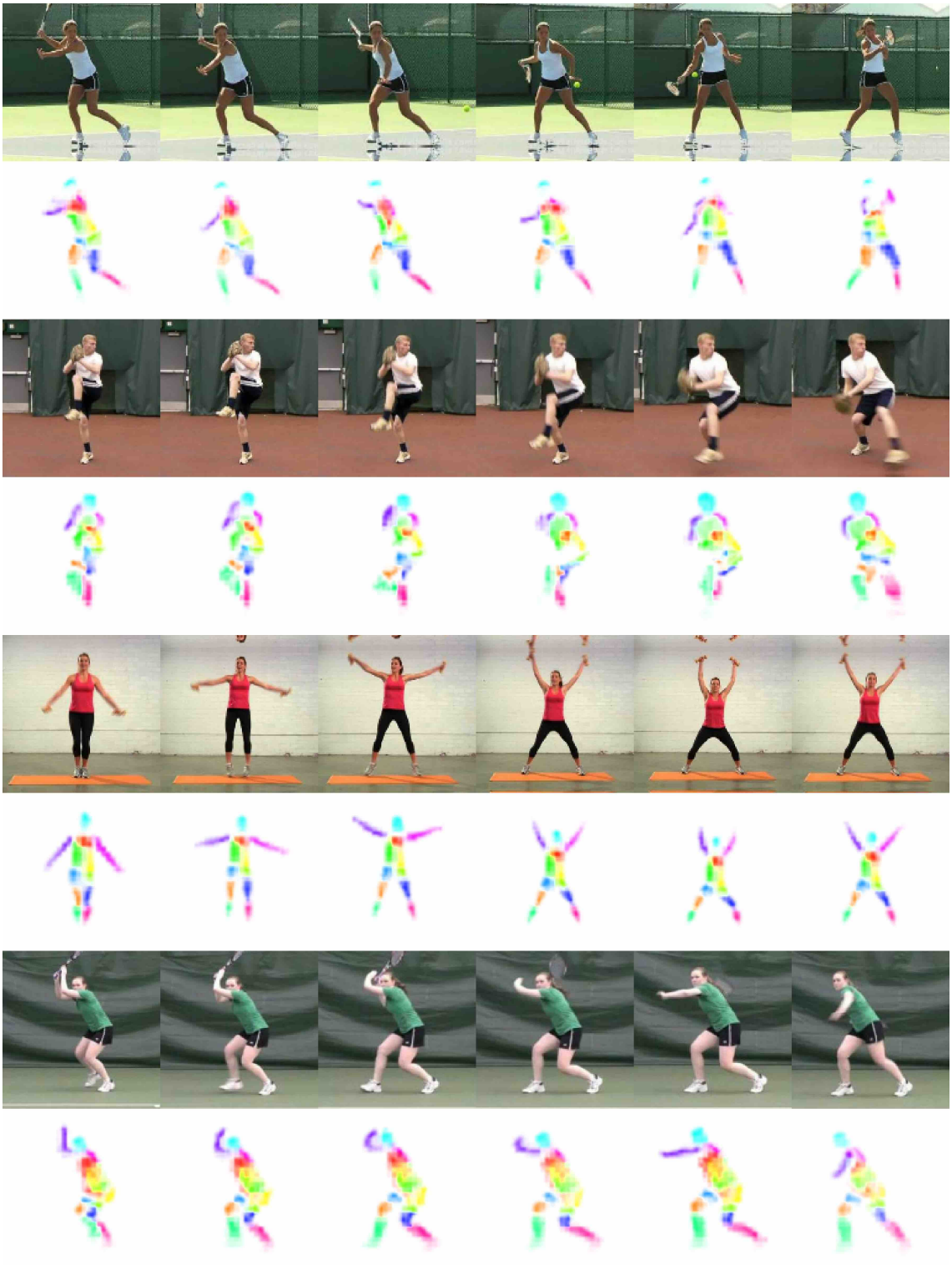


Figure A.1: Showing 12 out of 16 part activation maps on Penn Action.

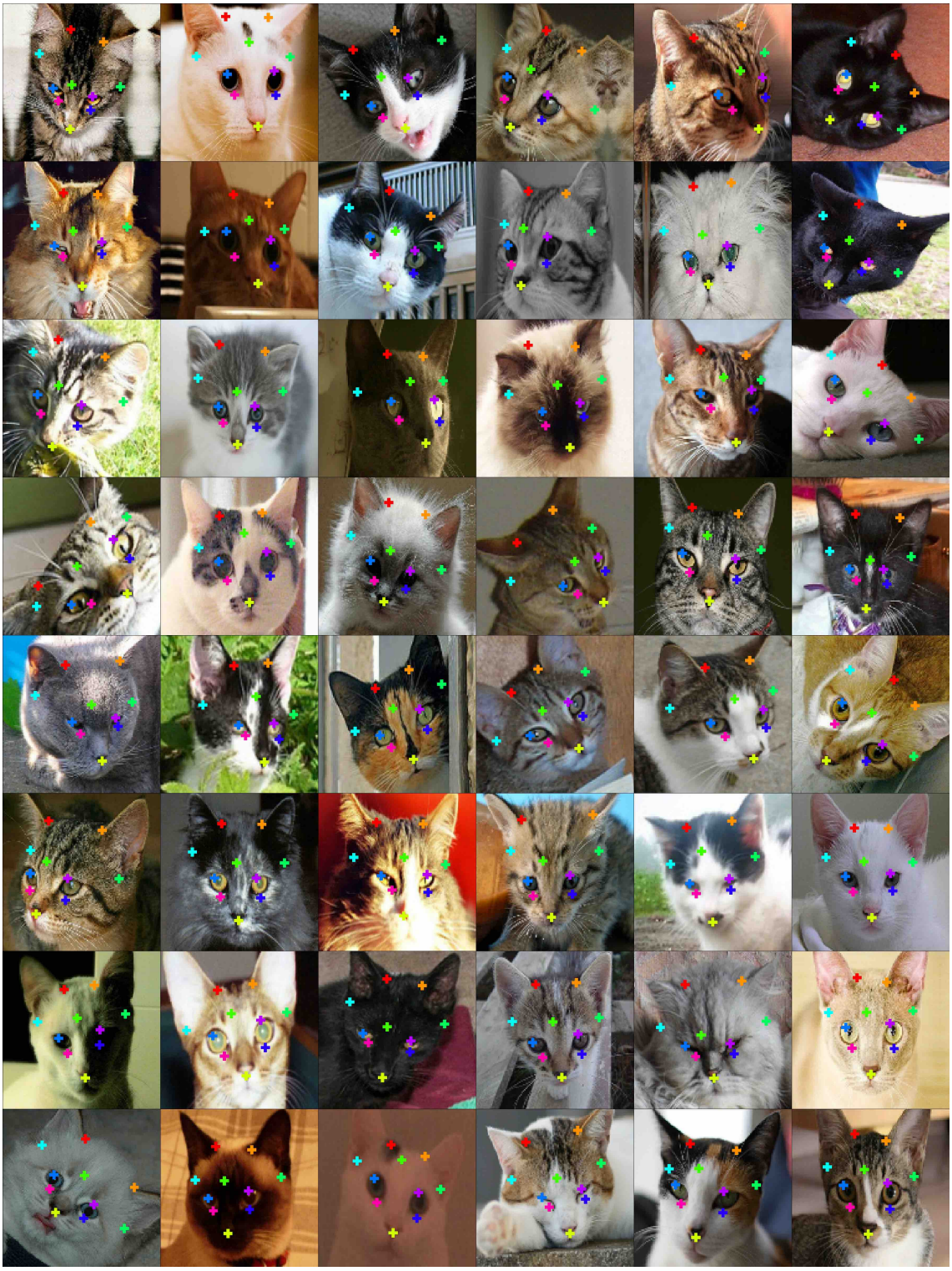


Figure A.2: Discovering 10 landmarks on Cat Head.

Figure A.3: Discovering 10 landmarks on Dogs Run.

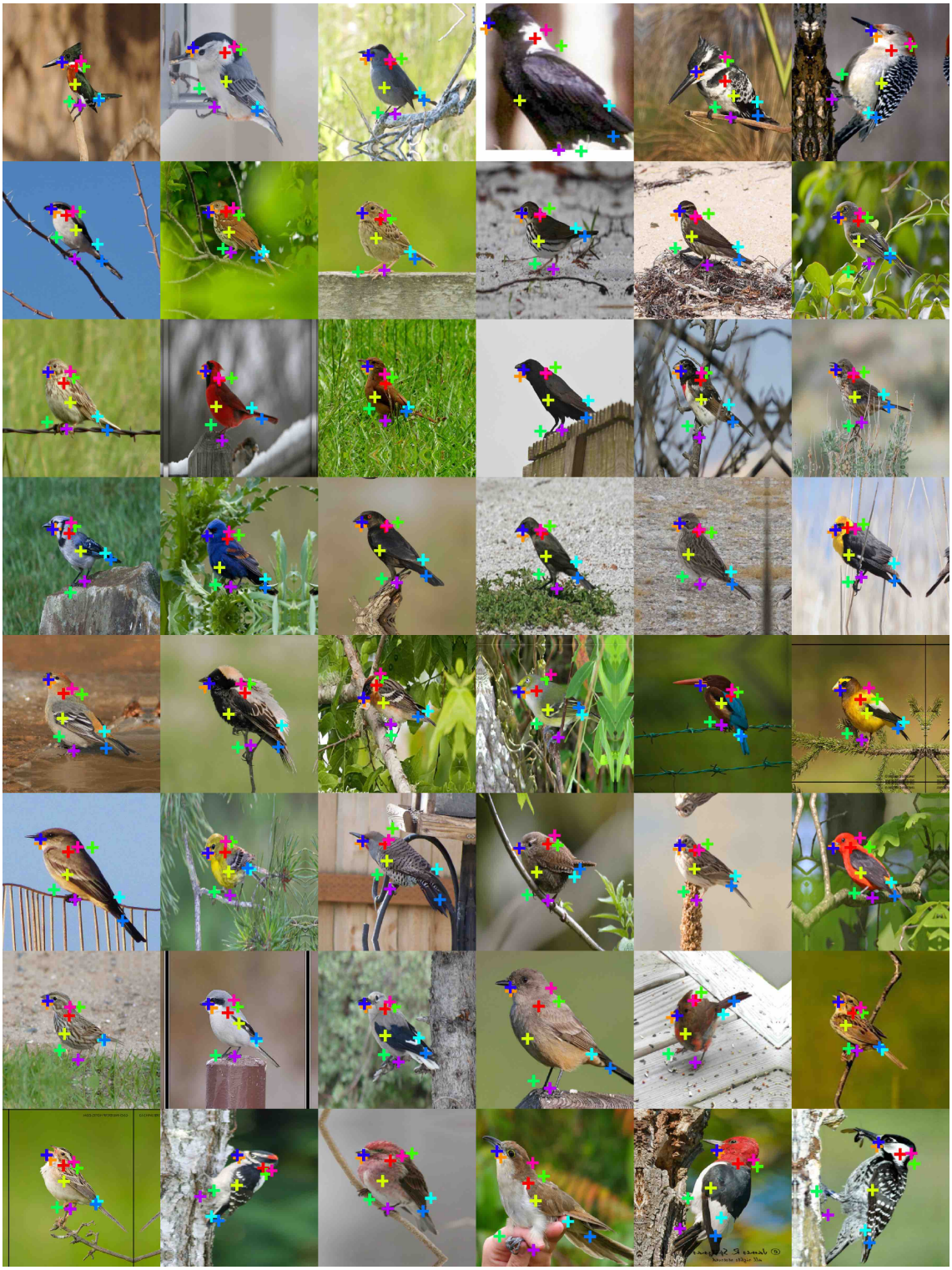


Figure A.4: Discovering 10 landmarks on CUB-200-2011.

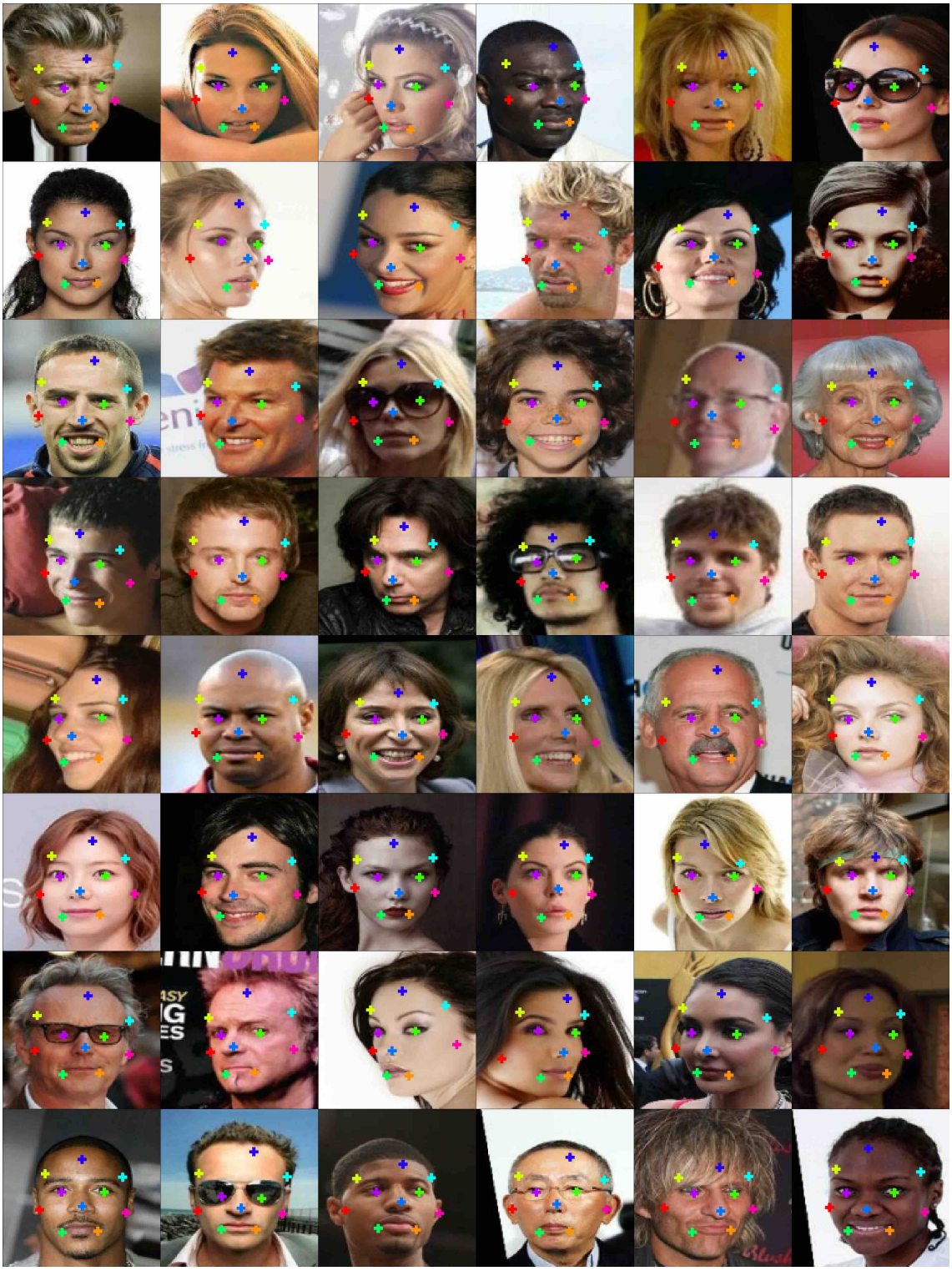


Figure A.5: Discovering 10 landmarks on CelebA.

Figure A.6: Discovering 10 landmarks on Human3.6M.

Figure A.7: Showing 12 out of 16 landmarks on Penn Action.

B Disentangled Representation

Local Appearance Transfer. In Fig. B.1, we show results for successively swapping part appearance on the Deep Fashion dataset.

Video-to-Video Translation. In Fig. B.2 we show sequences of a frame-to-frame appearance-shape transfer on the BBC Pose dataset. Note that the out-of-plane rotations and fine-grained details of hands and facial expressions are accurately captured. Notice the quality (smoothness, consistency) of the transfer.

Figure B.1: Successively altering the appearance of individual parts. We show 6 examples of successively altering appearances of parts using different source images. In each example we start from the original appearance (left-most column). The top row shows ground-truth images (taken from the test-set), which act as the source for the part appearance to be altered. The bottom row then illustrates the new synthesized image, which is generated based on the already altered part appearances plus the current appearance modification. Part appearances are altered in fixed order: head, upper body, legs, feet.

Figure B.2: Generated sequence on BBC Pose from a target pose sequence (leftmost column) and target appearances (top row).

C Implementation Details

Dataset	# landmarks	res.	lr.	advers.
Cat Head [73]	10 / 20	128×128	0.001	✗
CelebA [74]	10	128×128	0.001	✗
Human3.6M [76]	16	128×128	0.0002	✗
Penn Action [77]	16	128×128	0.0002	✗
Dogs Run (own)	12	128×128	0.001	✗
CUB-200-2011 [78]	10	128×128	0.001	✗
BBC Pose Regression [75]	30	128×128	0.001	✗
BBC Pose Synthesis [75]	40	256×256	0.001	✓
Deep Fashion [80, 81]	16	256×256	0.001	✓

Table C.1: Settings for different experiments: number of landmarks, input resolution, learning rate of Adam optimizer, adversarial task

Implementation Details. Table C.1 gives an overview over the different settings for the datasets we used in our experiments.

The architecture of the encoder for shape E_σ and appearance E_α is based on the implementation of the stacked hourglass architecture [31]. In a first step the image with input resolution $h \times w \times 3$ is processed by a series of convolutions to image features of dimension $64 \times 64 \times 256$. The hourglass modules of E_σ and E_α operate on a maximal resolution of 64×64, thus part activation maps and the localized image appearance encoding both have a spatial dimension of 64×64. E_σ reaches its lowest resolution at 4×4 pixels whereas E_α has its lowest resolution at 32×32 pixels. All residual blocks of the hourglass modules have 256 feature channels. The decoder is a variant of a U-Net [32] operating at a resolution of $h \times w$ pixels. Different from a standard U-Net we do not learn the downsampling stream. Through skip connections the approximate part activations maps are passed to the up-sampling stream with the appropriate resolutions. We distribute the local appearance encoding together with the corresponding approximate part activation maps into a multi-scale bottleneck of resolution 4×4 to 16×16 in the U-Net. The convolutional filters in the first up-sampling stage of the U-Net have 512 feature channels. The number of feature channels is halved every two up-sampling stages.

Local Loss. The ℓ_1 reconstruction loss is weighted locally around the part activations $\sigma^i(\mathbf{x})$. For this, we multiply the loss with a soft mask. For an image $\mathbf{x}$ at pixel u the mask

takes the form:

$$\text{mask}[u] = \min\Big(\sum_i \frac{1}{1 + \|u - \mu[\sigma^i(\mathbf{x})]/\lambda_{\text{scal}}\|}, 1\Big), \tag{C.1}$$

where λ_{scal} is a hyperparameter. We do not propagate gradients through the means $\mu([\sigma^i(\mathbf{x})])$ of the mask.

Adversarial Task. To improve the quality of image generations, we implement a variant of the adversarial task, as presented in [22]: A discriminator is trained to classify $N \times N$ image patches as real or fake. Using the mean locations of part shapes as center points, we extract image patches of size 49×49 from the real image $\mathbf{x}$ and the generated image $\hat{\mathbf{x}}$. As conditioning, the discriminator is additionally provided with corresponding patches extracted on the stack of approximated part activations $\tilde{\sigma}^i(\mathbf{x})$. The discriminator is implemented as a lightweight CNN architecture consisting of 4 convolution layers with stride 2 followed by a dense layer. The adversarial task is trained simultaneously with the main objective function, no subsequent fine-tuning step is necessary.

D Dataset Preprocessing

CelebA

CelebA [74] contains ca. 200k celebrity faces of 10k identities. We resize all images to 128×128 and exclude the training and test set of the MAFL subset, following [29]. As [29, 30], we train the regression (to 5 ground truth landmarks) on the MAFL training set (19k images) and test on the MAFL test set (1k images).

Cat Head

Cat Head [73] has nearly 9k images of cat heads. We use the train-test split of [30] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [30]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.

CUB-200-2011

CUB-200-2011 [78] comprises ca. 12k images of birds in the wild from 200 bird species. We excluded bird species of seabirds, roughly cropped using the provided landmarks as bounding box information and resized to 128×128. We aligned the parity with the information about the visibility of the eye landmark. For comparing with [30] we used their published code.

BBC Pose

BBC Pose [75] contains videos of sign-language signers with varied appearance in front of a changing background. Like [65] we loosely crop around the signers. The test set includes 1000 frames and the test set signers did not appear in the train set. For evaluation, as [65], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M

Human3.6M [76] features human activity videos. We adopt the training and evaluation procedure of [30]. For proper comparison to [30] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset.

Penn Action

Penn Action [77] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

Dogs Run

Dogs Run is made from dog videos from YouTube totaling in 1250 images under similar conditions as in Penn Action. The dogs are running in one direction in front of varying backgrounds. The 17 different dog breeds exhibit widely varying appearances.

Deep Fashion

Deep Fashion [80, 81] consists of ca. 53k in-shop clothes images in high-resolution of 256×256. We selected the images which are showing a full body (all keypoints visible, measured with the pose estimator by [62]) and used the provided train-test split. For comparison with Esser *et al*. [66] we used their published code.

E Lists

E.1 List of Figures

E.2 List of Tables

F Bibliography

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 7, 16, 26

[2] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM*, 2018. 7, 10, 17, 18, 69

[3] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 7, 15

[4] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 7, 15, 26, 50

[5] Sridhar Mahadevan. Imagination machines: A new challenge for artificial intelligence. In *AAAI*, 2018. 7

[6] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 8, 67

[7] Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018. 8, 9, 13, 19, 67

[8] Carl G Jung. Collected works of cg jung: The archetypes and the collective unconscious (vol. ix), 1968. 9

[9] Noam Chomsky et al. *New horizons in the study of language and mind*. Cambridge University Press, 2000. 9

[10] Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B Tenenbaum, and Luca L Bonatti. Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, 2011. 9

[11] Matthew B Wall and Andrew T Smith. The representation of egomotion in the human brain. *Current Biology*, 2008. 9

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *ICCV*, 2015. 10, 45

[13] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Güler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*, 2018. 10, 28

[14] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *ECCV*, 1998. 10, 27

[15] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 12, 26

[16] Christopher M Bishop. Pattern recognition and machine learning (information science and statistics). 2006. 12, 15

[17] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv*, 2015. 13

[18] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 1989. 14

[19] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1991. 14

[20] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 14

[21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv*, 2015. 16

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv*, 2017. 16, 64

[23] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017. 18

[24] H. Reichenbach. *The Direction of Time*. University of California Press, 1956. 18

[25] Judea Pearl and Dana Mackenzie. *The Book of Why*. Hachette Book Group, 2018. 18

[26] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv*, 2018. 19, 26

[27] Tijmen Tieleman. *Optimizing neural networks that generate images*. University of Toronto (Canada), 2014. 20, 26, 50

[28] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv*, 2012. 20, 26

[29] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 22, 27, 29, 31, 33, 34, 38, 65

[30] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 22, 27, 31, 32, 33, 34, 35, 36, 37, 38, 40, 43, 65, 68

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 23, 27, 34, 38, 63

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 24, 63

[33] Thomas G Bever and David Poeppel. Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics*. 26

[34] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*. 26, 50

[35] Ilker Yildirim, Tejas D Kulkarni, Winrich Freiwald, and Joshua B Tenenbaum. Efficient analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *CogSci*, 2015. 26

[36] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *NIPS*, 2016. 26

[37] 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. *ICLR*, 2017. 26

[38] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. *ICLR*, 2018. 26

[39] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv*, 2018. 26

[40] Zejian Li, Yongchuan Tang, and Yongxing He. Unsupervised disentangled representation learning with analogical relations. In *IJCAI*, 2018. 26

[41] David A Ross and Richard S Zemel. Learning parts-based representations of data. *JMLR*, 2006. 27, 37

[42] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychol. Rev.*, 1987. 27

[43] Pedro F Felzenszwalb, Ross B Girshick, David A McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 27

[44] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017. 27

[45] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 27

[46] Grégoire Mesnil, Antoine Bordes, Jason Weston, Gal Chechik, and Yoshua Bengio. Learning semantic representations of objects and their parts. *Mach Learn*, 2013. 27

[47] Wei Yang, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 27

[48] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2017. 27

[49] Tu Dinh Nguyen, Truyen Tran, Dinh Q Phung, and Svetha Venkatesh. Learning parts-based representations with nonnegative restricted boltzmann machine. In *ACML*, 2013. 27

[50] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. *CVPR*, 2015. 27

[51] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *TPAMI*, 2017. 27

[52] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 27

[53] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 2016. 27

[54] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 27

[55] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 27

[56] Marco Pedersoli, Radu Timofte, Tinne Tuytelaars, and Luc J Van Gool. Using a deformation field model for localizing faces and facial points under weak supervision. In *CVPR*, 2014. 27

[57] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *ICCV*, 2011. 27

[58] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 27

[59] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 27, 34, 39

[60] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 27

[61] Jongin Lim, Youngjoon Yoo, Byeongho Heo, and Jin Young Choi. Pose transforming network: Learning to disentangle human posture in variational auto-encoded latent space. *Pattern Recognit. Lett.*, 2018. 27

[62] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 27, 45, 66

[63] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCV Workshops*, 2016. 27

[64] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017. 27

[65] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 27, 31, 32, 34, 37, 38, 39, 44, 65, 68

[66] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 28, 44, 45, 46, 66, 70

[67] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 28, 44

[68] Rodrigo de Bem, Arnab Ghosh, Thalaiyasingam Ajanthan, Ondrej Miksik, N Siddharth, and Philip H S Torr. Dgpose: Disentangled semi-supervised deep generative models for human body analysis. *arXiv*, 2018. 28, 44, 48

[69] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017. 28, 44

[70] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. *CVPR*, 2018. 28

[71] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John V Guttag. Synthesizing images of humans in unseen poses. *arXiv*, 2018. 28

[72] Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv*, 2018. 28

[73] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 31, 63, 65

[74] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 31, 63, 65

[75] James Charles, Tomas Pfister, Derek R Magee, David C Hogg, and Andrew Zisserman. Domain adaptation for upper body pose tracking in signed tv broadcasts. In *BMVC*, 2013. 32, 34, 39, 63, 65

[76] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2014. 33, 63, 65

[77] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 33, 63, 66

[78] C Wah, S Branson, P Welinder, P Perona, and S Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 34, 63, 65

[79] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 44

[80] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 44, 45, 63, 66

[81] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 44, 45, 63, 66

[82] Jon Almazán, Bojana Gajic, Naila Murray, and Diane Larlus. Re-ID done right: towards good practices for person re-identification. *arXiv*, 2018. 45

[83] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image Vision Comput.*, 2014. 45

[84] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person Re-identification: Past, Present and Future. *arXiv*, 2016. 45

[85] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 45

[86] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 46

[87] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 46

[88] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An Image-based virtual try-on network. *arXiv*, 2017. 48

[89] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A Neural Algorithm of Artistic Style. *arXiv*, 2015. 50

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den (Datum) .....