Contents

1 Introduction

Computer vision is the scientific endeavour to algorithmically understand patterns in images. Structures and processes 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 use this mirror as a window to observe the reality lurking behind it, *i.e.* to measure the causal elements that contributed to the image generation. Typically, objects appear in an intricated interaction of many factors of variation. For example, given the object class of people, persons can vary in their visual appearance by clothing and skin color or in their geometric structure due to their pose or body physique. For articulated object classes the most prominent factors of variation are geometric shape and visual appearance. Disentangling these factors is a difficult problem, due to the intricated interplay of shape and appearance, especially under heavy articulation. The complexity enters, as a variation in shape is a change of the images domain rather than a change of its values [1]. Consider a person raising his arm: the color and texture of his pullover sleeve intrinsically does not change, but appears at a different location in the image. An efficient model for shape should cover all possible states of the object and preserve the local linkage to its intrinsic appearance.

1.1 Why Disentangle Causal Factors?

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 [2]. 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, humans may change images at will, but also machines may reason about the world [3], 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"* are manageable tasks for the human imagination, but are out of the league for currently used generative image models [4, 5], 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 [6]. 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.

Figure 1.1: The image captions are generated by a deep neural network (Neuraltalk2) [7]. Yet, common sense understanding of psychological and physical entities in terms of causal relationships and narratives is absent [8]. Instead, the neural network seems to capture mere associations.

1.2 How not to Disentangle.

Can machines tell a story? Carefully observe your own mind, when viewing the images shown in Fig. 1.1: 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 [7]), 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 , disentangling causal factors from raw image data, without any side information is impossible theoretically, and can only work based on statistical assumptions. Lets consider an example: 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)? Lets suppose the distribution of variation follows a two-dimensional Gaussian distribution, one dimension for pose, one for appearance. The learning algorithm has access to randomly sampled images from this distribution. An intelligent 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 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 judgement) 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 moving without loosing clothes is a trivial example for that. In summary, on the basis of dataset statistics 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, since 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. 1.1). 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 [8]. Archetypal knowledge of psychology [9], a universal grammar for language [10] and causal intuitions on everyday physics [11] 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 change, not only by observing change (as in a temporal sequence), but also by changing. 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. *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 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 *data* that contains the necessary causal relationships and interactions, to learn a useful *model* of the world.

1.4 Contributions

This thesis makes two theses:

- *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 [3], disentangling causal factors requires model assumptions and/or interactional 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*, we give an overview over state-of-the-art unsupervised disentangling literature and situate the proposed method in relation to the literature. 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.

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. 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.

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 [13]. 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 $x_i \sim p(x)$.

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 *e.g.* 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. 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 often arbitrary and subjective.

- 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 [8]. A related modeling choice is, if the algorithm should be model-free or model-based.

2.1.3 Artificial Neural Networks

Artificial neural networks are a powerful and flexible tool for function approximation. In their principles they are inspired by biological neurophysiology. An artificial network is a model for a function $y = f(x)$ with vector input $x = \{x_i | i = 1 \dots n\}$ and vector output $y = \{y_j | j = 1 \dots m\}$:

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

, with weight matrices w, w', non-linear so-called activation functions a, a' (*e.g.* $a(x) = 0$ for $x < 0$, $a(x) = x$ for $x >= 0$) and bias vectors b, b'. The components h_j are called hidden units or neurons. Neural networks can also comprise multiple hidden layers via $h_j = a(\sum_i w_{ji} h_i + b_i)$. It can be shown, that in the limit of infinite hidden units h_j they can approximate any (continuous) function arbitrarily close [14, 15]. In practice, however, networks with more that one layer, referred to as deep neural networks, seem to work better. This may be due to the possibility of building a hierarchical feature representation [16].

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.

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 $\hat{p}(x)$. Thus, after estimation, one can generate samples from this model $\hat{p}$. Hence the name generative model. 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* 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:

$$\mathcal{L} = \mathbb{E}_{z\sim q(z|x)} \log p(x|z) - \mathrm{KL}(q(z|x)||p(z)) \tag{2.2}$$

Where z introduces latent variables, with a prior distribution $p(z)$. The 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 or moments (μ, Σ) are modelled with neural networks.

2.2.2 Adversarial Formulations

Generative adversarial networks (GAN) [4] 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:

$$\min_D \max_G -\frac{1}{2}\mathbb{E}_{x\sim p_{data}}[\log D(x)] - \frac{1}{2}\mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))] \tag{2.3}$$

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 [17]. There are many variants and extensions to this basic principle of learning with an adversarial task. For example, one can learn a discriminator on for a set of image patches [18].

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*. [2]

According to Bengio *et al*. [2], 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 by Equivariance and Invariance

What is a factor? As outlined in the introduction (cf. sec. **??**), factors in a representation should correspond to 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. **??**). 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. Formally, this can be seen as inference: a number of latent variables $\mathbf{z_1} \dots \mathbf{z_N}$ interacted to cause the existence of the observed image $\mathbf{x}$. The task is now to infer estimates for these latent variables $\hat{\mathbf{z}}_\mathbf{i}$. A graphical model of the process is shown in figure 2.1. A disentangled representation should simultaneously fulfill equivariance and invariance: A change in $\mathbf{z_i}$ should: *i) equivariantly* change in the abstract representational factor $\hat{\mathbf{z}}_\mathbf{i}$, *ii)* while leaving the other factors $\hat{\mathbf{z}}_\mathbf{j}, j \neq i$, that represent other causes, *invariant*.

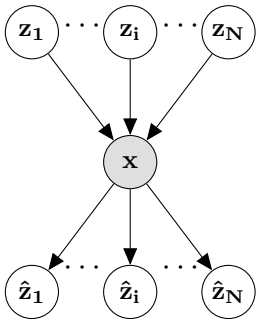


Figure 2.1: Disentangling causal factors means to infer an estimate - *i.e*. a representation - from an image

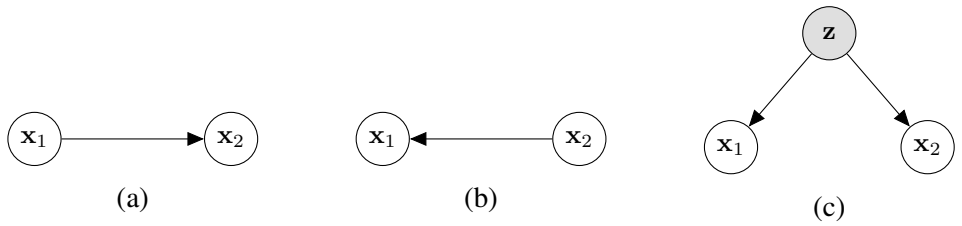


Figure 2.2: Correlation implies causation - if x_1 and x_2 correlate, a) x_1 may cause x_2, b) x_1 may be caused by x_2 or c) both are contingent on a latent cause z

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, 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 - raw data meaning images x sampled from $p(x)$, without further assumptions, it is impossible learn a disentangled representation z. We start with a primer for causal learning (sec. 2.4.1), outline which inductive biases are needed for disentanglement (sec. 2.4.2) and assess how one can instantiate such biases for disentangling the factors of shape and appearance in images (sec. 2.4.3, sec. 2.4.4)).

2.4.1 Causal Learning

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. A well-known theme is that statistical correlation does not imply causation. Less well-known is Reichenbachs principle [19, 20], 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. 2.2). In addition to estimating the probability distribution, also the causal structure has to be inferred [19].

We start with an intuitive example problem: 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 humidity to needle position. For example a model of: weather influencing air pressure influencing barometer needle position. What if one has no prior knowledge? A solution of childlevel simplicity is, to force the needle to move with a finger. Without the power of voodoo 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 [21] distinguishes between three types of questions, that can be answered by different types of knowledge:

1. Association. What if I see . . . ?
2. Intervention. What if I do . . . ?
3. Counterfactual. What if I had done . . . ?

The levels of this *ladder of causation* [21] 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 rigerous proof for this can be found in [22]. 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. 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.

2.4.3 Image Transformation as Intervention

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.

2.4.4 Analysis-by-Synthesis to Model Assumptions

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 that estimating the factors from

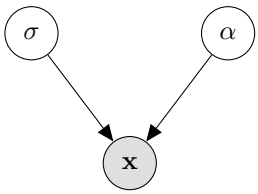


Figure 2.3: An image $\mathbf{x}$ is assumed to be generated from the factors of shape σ and appearance α. Implementing an intervention with a transformation of factors, means changeing one factor without changing the other.

Figure 2.4

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 [23]. Assumptions about how shape and appearance interact enable disentanglement.

3 Analysis of Literature on Disentangling

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 unsupervised shape learning.

3.1 Analysis-by-Synthesis

Analysis-by-synthesis is a theme that originates from the research on language perception *e.g.* [24], but has also been successfully applied to visual perception. The idea is, to guide the perception (analysis) by a model for generation (synthesis). In computational vision this can be formulated in autoencoder models [23]. In vision, the domain-specific knowledge about the image generation (decoding) exceeds the knowledge about computational cognition (encoding). 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 [25, 26, 27], 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.

3.2 Disentangled Generative Models

In order to gain a conceptual understanding of the world, disentangling the underlying factors of variation is a crucial step, as has been argued in numerous works, [27, 2, 28, 29, 30]. 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 [13] and autoencoding [5] model formulations. Recently, the endeavour for disentangling explanatory factors in the latent representation of generative models is being made explicit in the objective functions [31, 28] of these models. So far, however, these attempts are limited to rigid objects without articulation and disentangle holistic image factors like illumination, object rotation or total shape and global appearance [32].

3.3 Part-based Representation Learning

Describing an object as an assembly of parts is a classical paradigm for learning an object representation in computer vision [33] with linkage to human perceptual theories [34]. 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 [35, 36, 37, 38, 39, 40]. 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 [33, 41, 42].

3.4 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. Landmark learning has been applied to human faces [43, 44, 45, 46, 47, 48, 49] and bodies [50, 51, 52, 53, 54, 55, 56].

To tackle the problem without supervision, Thewlis *et al*. [57] proposed enforcing equivariance of landmark locations under artificial transformations of images. The equivariance idea had been formulated in earlier work [58] and has since been extended to learn a dense object-centric coordinate frame [59]. However, enforcing only equivariance encourages consistent landmarks at discriminable object locations, but disregards an explanatory coverage of the object.

Zhang *et al*. [60] addresses this issue: 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.

Jakab *et al*. [61] 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. Thus, parts are learned which meaningfully capture the variance of an object class in shape as well as in appearance.

Additionally, and in contrast to all these works ([57, 60, 61]) we take the extend of parts into account, when formulating our equivariance constraint. Furthermore, we ex-

plicitly address the goal of disentangling shape and appearance on a part-based level by introducing invariance constraints.

3.5 Disentangling Shape and Appearance

Factorizing an object representation into shape and appearance is a popular ansatz for representation learning. Recently, a lot of progress has been made in this direction by conditioning generative models on shape information [62, 63, 64, 65, 66, 67]. While most of them explain the object holistically, only few also introduce a factorization into parts [66, 67]. In contrast to these shape-supervised approaches, we learn both shape and appearance without any supervision.

For unsupervised disentangling, several generative frameworks have been proposed [29, 28, 68, 32, 1, 69]. However, these works use holistic models and show results on rather rigid objects and simple datasets, while we explicitly tackle strong articulation with a part-based formulation.

4 Method

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, see Fig. 4.1. 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 auto-encoding 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. 4.1.
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. 4.2.

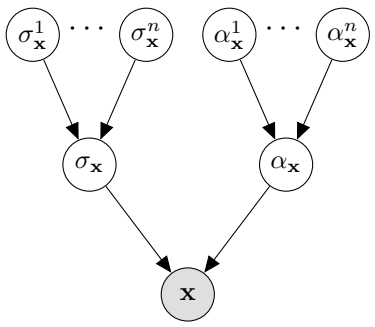


Figure 4.1: Modelling an image $\mathbf{x}$ of an object with shape $\sigma_{\mathbf{x}}$ and appearance $\alpha_{\mathbf{x}}$, by factorizing into part shapes $\sigma^i_{\mathbf{x}}$ and part appearances $\alpha^i_{\mathbf{x}}$

4.1 Transformation Framework

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

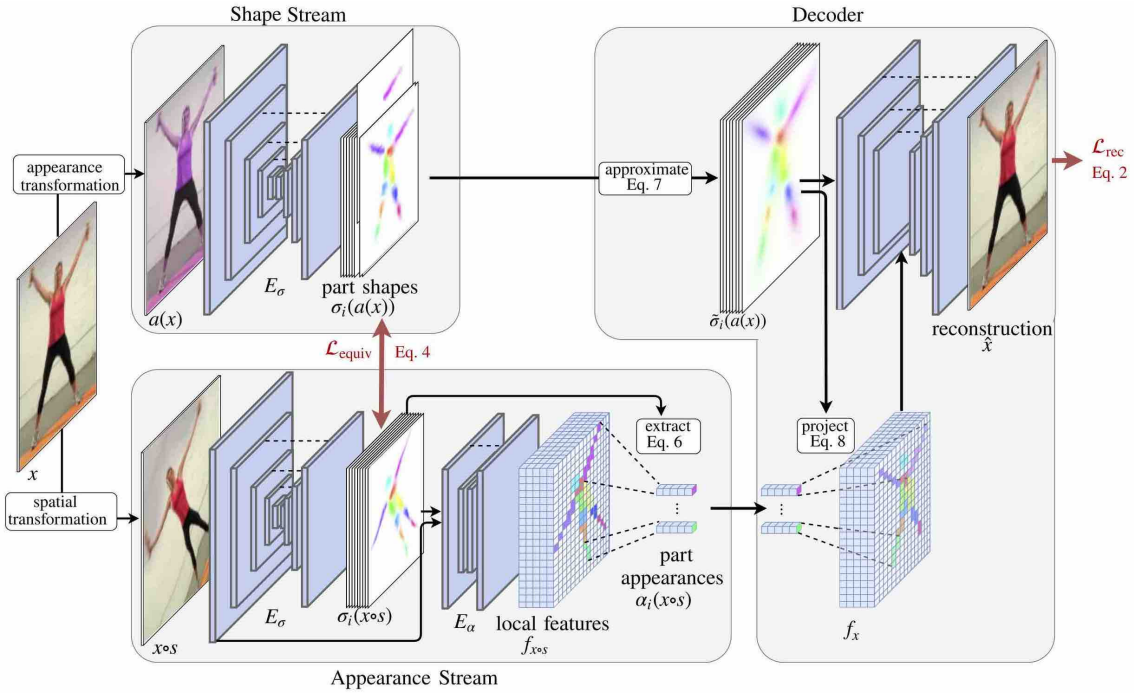


Figure 4.2: Encoder E encodes shape and appearance for two transformed images $s(\mathbf{x})$ and $a(\mathbf{x})$, after recombination R of $(\alpha_{s(\mathbf{x})}, \sigma_{a(\mathbf{x})})$ into latent image Z, the decoder D reconstructs the image $\mathbf{x}$.

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 auto-encoding paradigm, with part shapes and part appearances assuming the role of the latent code. To incorporate these constraints into the

loss of an auto-encoder, we reconstruct an image $\mathbf{x}$ not from the shape and appearance $(\alpha_{\mathbf{x}}, \sigma_{\mathbf{x}})$ 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. 4.2) to a latent image Z, a decoder D reconstructs the image. This configuration is depicted in Fig. 4.2, the reconstruction loss $\mathcal{L}_{\text{rec}}$ is as follows:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{x} - \mathrm{D}[\alpha_{s(\mathbf{x})}, \sigma_{a(\mathbf{x})}]\| \tag{4.1}$$

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. 4.2 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 [57, 60] 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 the mere position.

$$\mathcal{L}^i_{\text{equiv}} = \mathcal{L}^i_{\mu} + \mathcal{L}^i_{\Sigma} \tag{4.2}$$

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

$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}^i_{\text{equiv}} + \mathcal{L}_{\text{rec}} \tag{4.3}$$

4.2 Analysis-by-Synthesis Architecture

The auto-encoding pipeline consists of analysis and 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.

4.2.1 Analysis

The encoding of shape and appearance given an image $(\alpha, \sigma|\mathbf{x})$ 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 [54]. The network input is an image $\mathbf{x}$, the output a stack of n part shapes $s = \{\sigma^i | i = 1, ..., n\}$.
(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:

$$\alpha^i = \sum_{p \in \Lambda} A(p) \frac{\sigma^i(p)}{\sum_{p' \in \Lambda} \sigma i(p')}. \tag{4.4}$$

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

4.2.2 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 (cf. sec. 2.4.4).

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:

$$z^i(\mathbf{x}) = \sigma_i(\mathbf{x}) \cdot \alpha_i. \tag{4.5}$$

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

$$Z(x) = \sum_i \frac{z^i(\mathbf{x})}{1 + \sum_j z^j(\mathbf{x})}. \tag{4.6}$$

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 [70]. The latent image is scaled to different scales 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

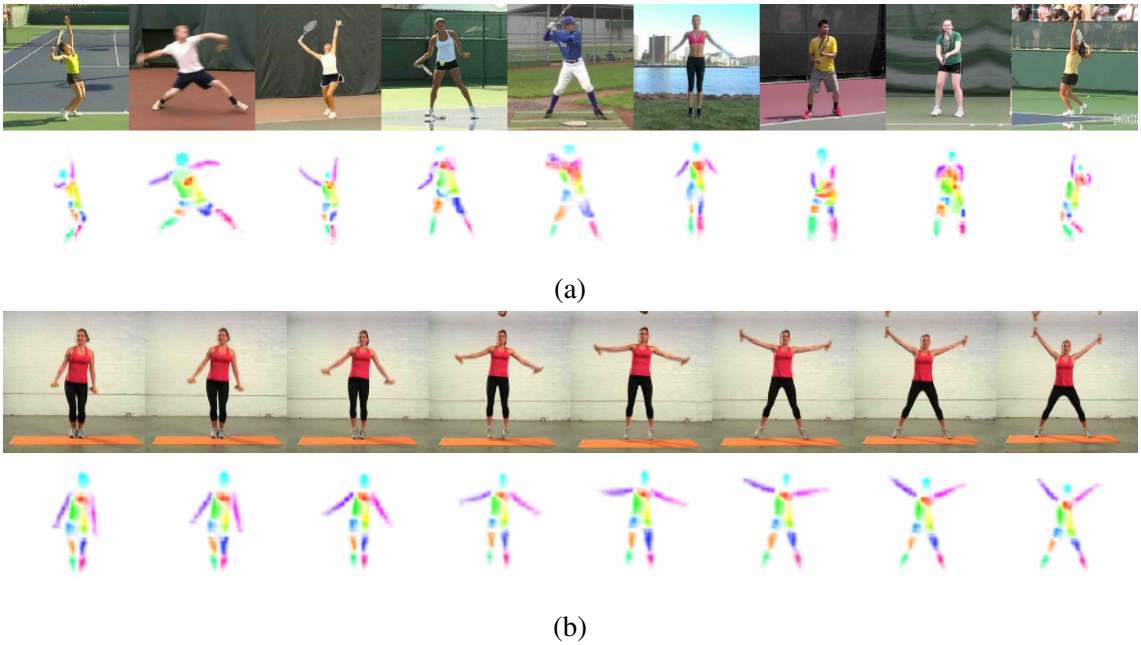
$$\hat{\sigma}_i(x) = \frac{1}{1 + (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)}, \tag{4.7}$$

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.

4.3 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 rescaled, 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. 4.7, 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.

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.

A visualization of the learned shape representation is shown in Fig. 5.1. 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*. [57], 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. On all datasets we outperform the state-of-the-art by a significant margin.

In the following we present first the quantitative and qualitative results by category, in

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 [57]	26.76	26.94	6.32
Jakab [61]	-	-	4.69
Zhang [60]	15.35	14.84	3.46
Ours	**9.88**	**9.30**	**3.24**

sec. 5.1, then highlight the challenges which occur (sec. 5.2) and argue for the importance of the transformations as a means to overcome those challenges (sec. **??**.

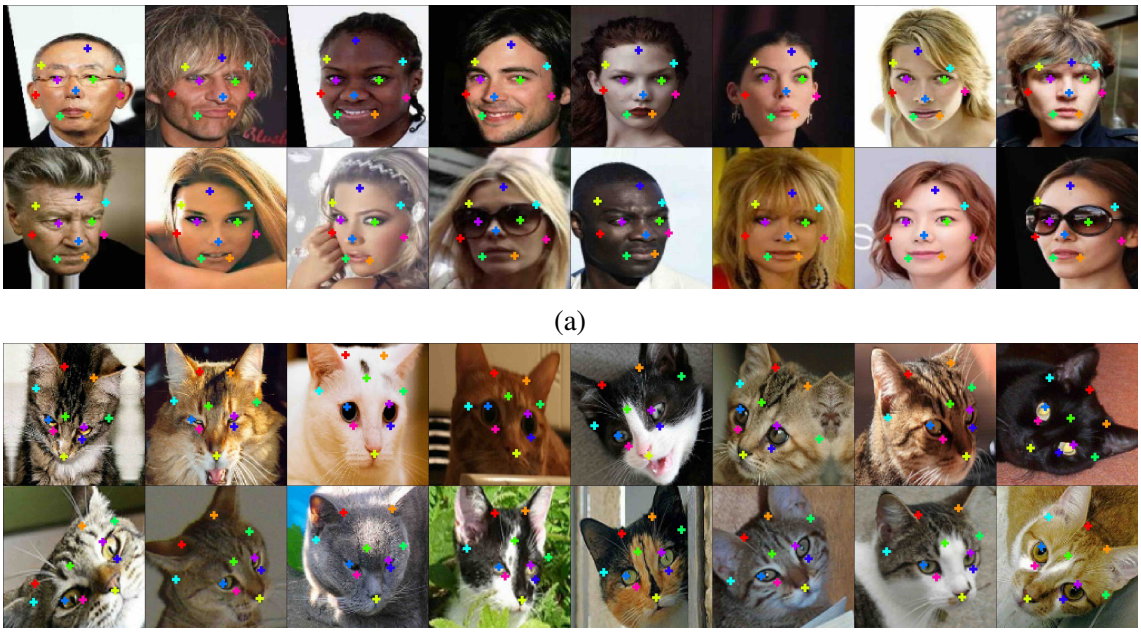
5.1 Results: Diverse Object Categories

On the object classes of human faces, cat faces, and birds (datasets CelebA, Cat Head, and CUB-200-2011) our model predicts landmarks consistently across different instances, cf. Fig. **??**. Tab. 5.4 compares against the state-of-the-art. Due to different breeds and species the Cat Head, CUB-200-2011 exhibit large variations between instances. Especially on these challenging datasets we outperform competing methods by a large margin. Fig. 5.6 also provides a direct visual comparison to [60] on CUB-200-2011. It becomes evident that our predicted landmarks track the object much more closely. In contrast, [60] 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.

5.1.1 Human and Cat Faces

CelebA [71] 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 [57]. As [57, 60], 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).

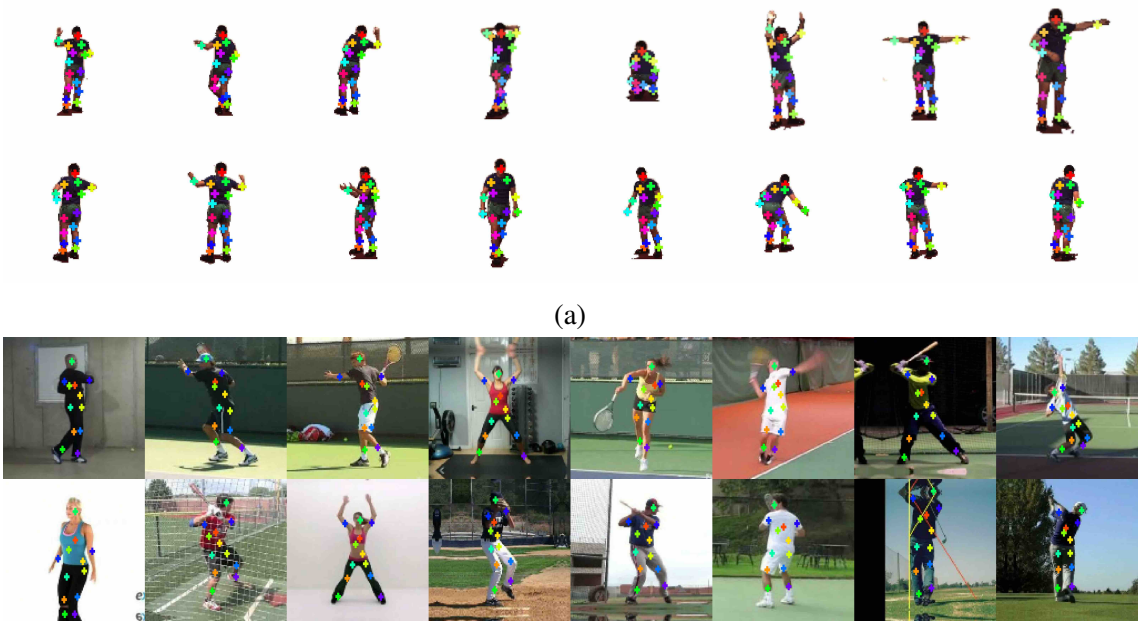
Cat Head [72] has nearly 9k images of cat heads. We use the train-test split of [60] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [60]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.



(a)

(b)

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



(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).

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 [73]	79.9%
	Pfister [52]	88.0%
unsupervised	Jakab [61]	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 [54]	2.16
semi-supervised	Zhang [60]	4.14
unsupervised	Thewlis [57]	7.51
	Zhang [60]	4.91
	Ours	**2.79**

5.1.2 Human Bodies

BBC Pose [73] contains videos of sign-language signers with varied appearance in front of a changing background. Like [61] 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 [61], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M [74] features human activity videos. We adopt the training and evaluation procedure of [60]. For proper comparison to [60] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset.

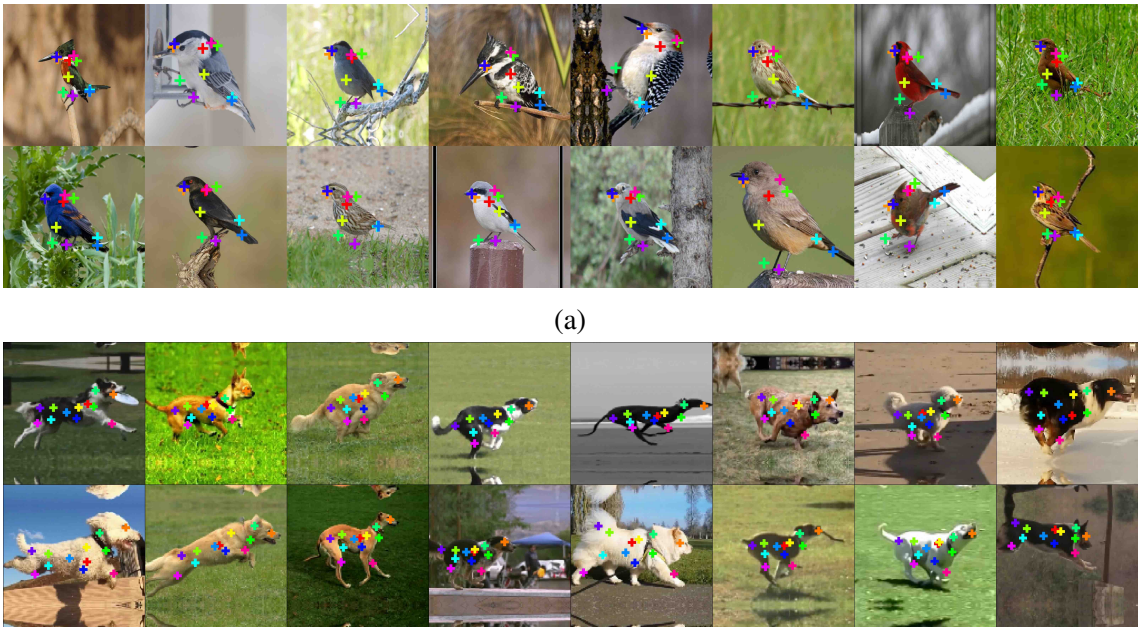
Penn Action [75] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

5.1.3 Animal Bodies

CUB-200-2011 [76] 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

Table 5.4: Error of unsupervised methods for landmark prediction on the CUB-200-2011 testing set. The error is in % of edge length of the image.

Dataset	CUB-200-2011
# Landmarks	10
Zhang [60]	5.36
Ours	**3.91**



(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).

the information about the visibility of the eye landmark. For comparing with [60] 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.

5.2 Challenges

An overview over the challenges implied by each dataset is given in Tab. 5.5.

We demonstrate (Fig. **??**), 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, which are present in

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

Dataset	Articul.	Var.	Backgr.	Viewp.
CelebA				
Cat Head		✓		
CUB-200-2011		✓	✓	
Human3.6M	✓			✓
BBC Pose	✓		✓	
Dogs Run	✓	✓	✓	
Penn Action	✓	✓	✓	✓

the Human3.6M or Penn Action datasets. Also with further complications such as viewpoint variations, blurred limbs and partial self-occlusions we are able to detect landmarks on Penn Action 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. The Dogs Run dataset displays that even completely different dog breeds can be related via semantic parts. The quantitative results 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, our approach is able to achieve a large performance gain even over results obtained with optical flow supervision. On BBC Pose, we outperform [**?**] by 6.1%, reducing the performance gap to supervised methods significantly.

5.2.1 Composite Objects/Scenes

What is an object? What is a scene? compositional nature of reality Bird on twig object? Bird can also fly, but neural networks learn by correlation in data (-> ref to these failure modes) Dancing pair as object.

5.2.2 Object/Background Separation

Complexly cluttered background - as long as no correlations to the object exist - is actually favorable for the method. Correlations of object with background will belong to object.

5.2.3 Object Articulation

Object articulation makes consistent landmark discovery challenging. Fig. **??** shows that our model exhibits strong landmark consistency under articulation and covers the full human body meaningfully. Even fine-grained parts such as the arms are tracked across heavy body articulations, which are frequent in the Human3.6M and Penn Action datasets. Despite further complications such as viewpoint variations or blurred limbs our model can detect landmarks on Penn Action of similar quality 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. Experiments on the Dogs Run dataset underlines that even completely dissimilar dog breeds can be related via semantic parts. Tab. 5.2 and Tab. 5.3 summarize the quantitative evaluations: we outperform other unsupervised and semi-supervised methods by a large margin on both datasets. On Human3.6M, our approach achieves a large performance gain even compared to methods that utilize optical flow supervision. On BBC Pose, we outperform [61] by 6.1%, reducing the performance gap to supervised methods significantly.

5.2.4 Intra-Class Variation

on cat head, human mislabelling is the most common failure mode

5.3 Transformations

Parity

birds parity salsa parity

Rotation, Scaling, Translation

on Cats -> black cats different set of KP than rest -> connect these samples via transformation to reach intra-class consistency

Mimicking Appearance

Figure 5.5: Examples for shape and appearance transformation on CUB-200-2011.

Color, Contrast, Hue

5.3.1 Natural Changes in Video Data

Video data: Penn, Own

Figure 5.6: Comparing discovered keypoints against [60] on CUB-200-2011. We improve on object coverage and landmark consistency. Note our flexible part placement compared to a rather rigid placement of [60] due to their part separation bias.

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

Table 5.6: Ablation studies on Cat Head dataset. We ablate the reconstruction loss $\mathcal{L}_{\text{rec}}$, equivariance loss $\mathcal{L}_{\text{eq}}$, the color augmentation and the crossing task

5.3.2 Ablating Contributions

We ablate the main components of our proposed framework: reconstruction loss $\mathcal{L}_{\text{rec}}$, the equivariance loss $\mathcal{L}_{\text{eq}}$, the appearance augmentation and the crossing task 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. 5.6 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 [60]. Training our model without color augmentations or appearance crossing between image pairs (i.e. the crossing task) weakens, respectively neglects the disentanglement of appearance and shape and hence the performance of our model significantly. Note that without the crossing task our models performs on par with Zhang *et al.* [60], indicating, that this novel task could be explaining overall performance gain w.r.t. [60]. Leaving out the explicit equivariance leads to the smallest drop in performance. This is not surprising, as equivariance is implicitly also enforced in the crossing framework.

5.4 Conclusion

We proposed a method to abstract away object appearance from shape. Despite the multifarious challenges in the diverse set of datasets the method is able to learn a dedicated part representation for shape. We compare the method to other approaches and reach a 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 [60] is the emphasis on disentanglement via a crossed reconstruction with shape and appearance transformations. Enforcing disentanglement via transformation 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.

Deep Fashion [77, 78] 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 [56]) and used the provided train-test split. For comparison with Esser *et al*. [62] we used their published code.

6 Disentangling Generative Factors

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 [62, 32, 63, 65, 64, 61], 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

On Deep Fashion [77, 78], a benchmark dataset for supervised disentangling methods, the task is to separate person ID (appearance) from body pose (shape) and then synthesize new images for previously unseen persons from the test set in eight different poses. We randomly sample the target pose and appearance conditioning from the test set. Fig. 6.1 shows qualitative results. We quantitatively compare against supervised state-of-the-art disentangling [62] by evaluating *i)* invariance of appearance against variation in shape by the re-identification error and *ii)* invariance of shape against variation in appearance by the distance in pose between generated and pose target image.

6.1.1 ReID

- t-SNE of IDs
- Own, Other (stronger statement)

To evaluate appearance we fine-tune an ImageNet-pretrained [12] Inception-Net [79] with a re-identification (ReID) algorithm [80] via a triplet loss [81] to the Deep Fashion

Table 6.1: 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 [62] is supervised w.r.t. shape.

	mAP	rank-1	rank-5	rank-10
VU-Net [62]	88.7%	87.5%	98.7%	99.5%
Ours	90.3%	89.4%	98.2%	99.2%

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.

training set. On the generated images we evaluate the standard metrics for ReID, mean average precision (mAP) and rank-1, -5, and -10 accuracy in Tab. 6.1. Although our approach is unsupervised it is competitive compared to the supervised VU-Net [62].

6.1.2 Pose

To evaluate shape, we extract keypoints using the pose estimator [56]. Tab. 6.2 reports the difference between generated and pose target in percentage of correct keypoints (PCK), Fig. 6.4 shows the comparison of PCK curves. As would be expected, VU-Net performs better, since it is trained with exactly the keypoints of [56]. Still our approach achieves an impressive PCK without supervision underlining the disentanglement of appearance and shape.

6.2 Factorizing into Parts

- Own Dataset: Move KP

Table 6.2: Percentage of Correct Keypoints (PCK) for pose estimation on shape/appearance swapped generations. α is pixel distance divided by image diagonal. Note that [62] serves as upper bound, as it uses the groundtruth shape estimates.

α	2.5%	5%	7.5%	10%
VU-Net [62]	95.2%	98.4%	98.9%	99.1%
Ours	85.6%	94.2%	96.5%	97.4%

Figure 6.2: Video-to-video translation on BBC Pose. Top-row: target appearances, left: target pose. Note that even fine details in shape are accurately captured. See supplementary for videos.

- DeepFashion: exchange parts

6.3 Follow-Up

- make generative:(KP distribution estimation, variational features).
- make video generation possible (RNN on KP vector).
- better transformations -> appearance locally (around parts changed), appearance changed perceptually -> style transfer

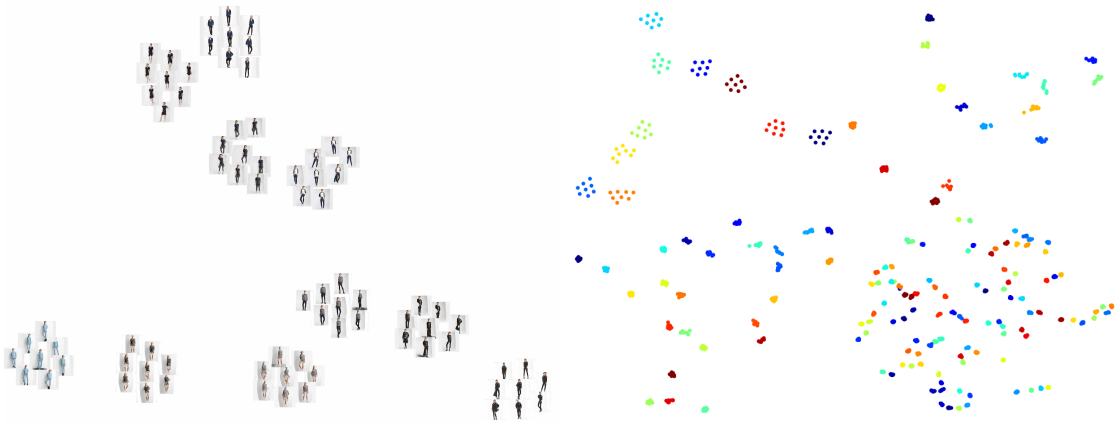


Figure 6.3: Visualization of feature distribution for generated person. (Right) t-SNE (perplexity 16) of 10 generated IDs, (left) color-coded t-SNE (perplexity 12) for 10, 15, 20 and 100 IDs. Each ID has 8 samples. The different IDs are clearly separable, despite variation in pose: Hence, generated appearance is invariant to pose.

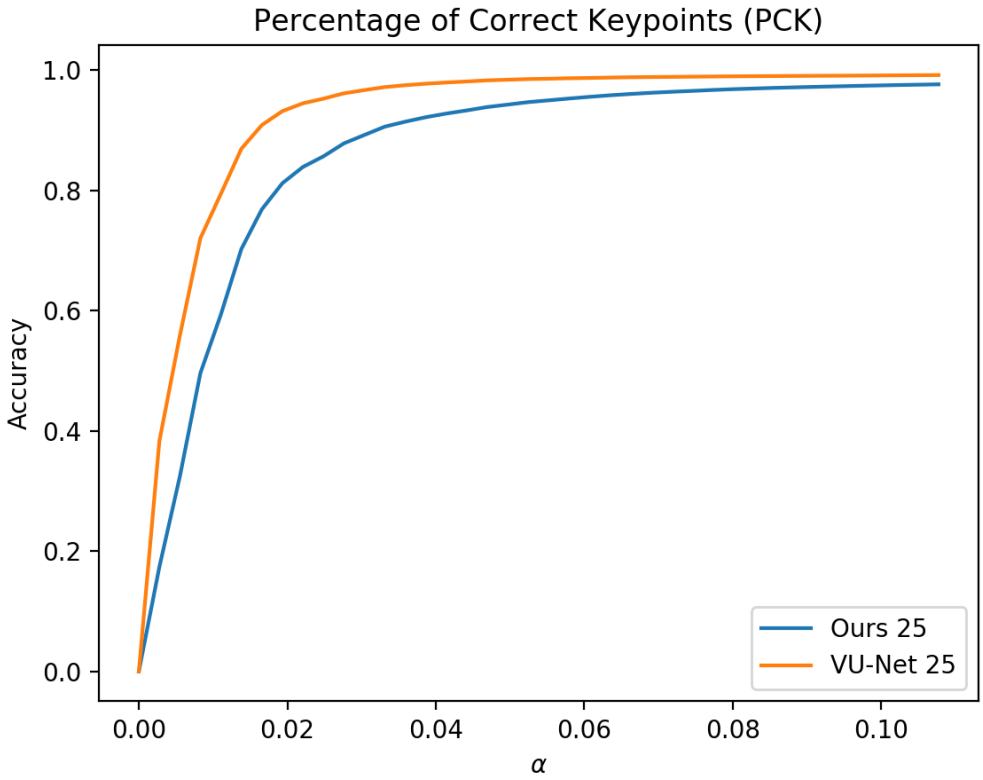
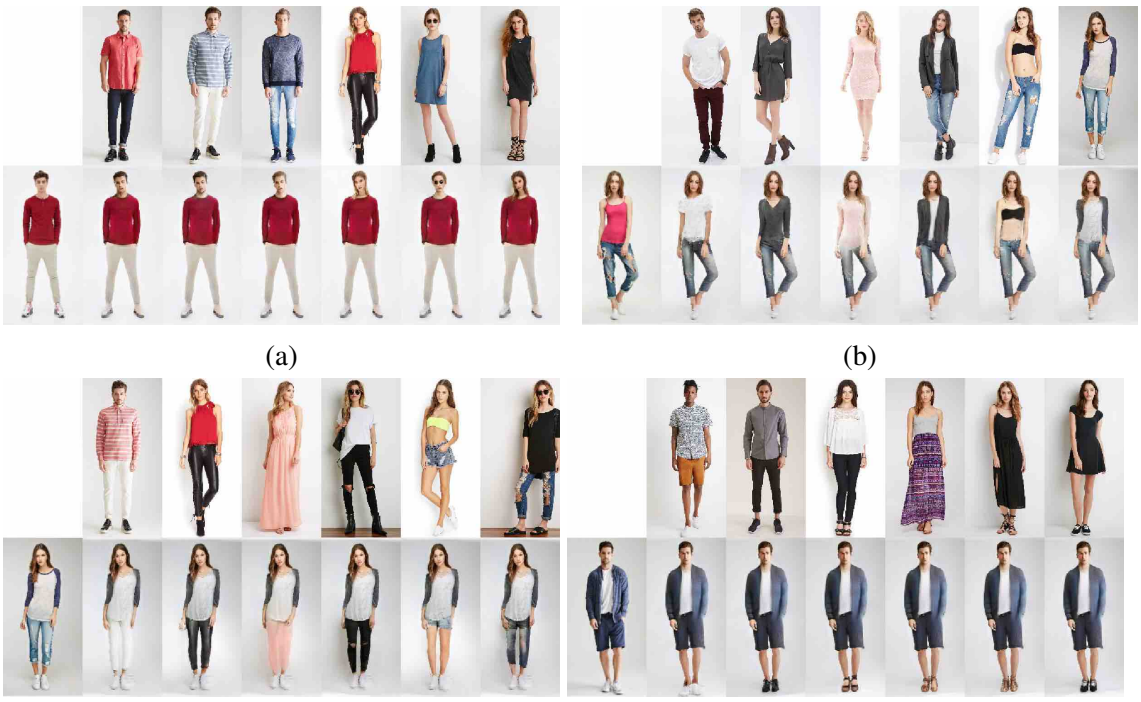


Figure 6.4: PCK Curve for VU-Net [62] and Ours for re-estimating pose with a 25 keypoint human pose detector.



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

Figure 6.5: 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.

7 Conclusion

-> need model-based approach (for counterfactual) make model as good as we can implementing as many assumptions as we can and only leave the rest to powerful model (humans also have brain structure and reasoning structure genetic)

need disentangling generative factors for imagination (i.e. synthesis) for manipulating factors mentally

8 Bibliography

[1] 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. 3, 16

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *TPAMI*, 2013. 3, 10, 14

[3] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM*, 2018. 3, 6

[4] 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. 3, 9

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 3, 14

[6] Sridhar Mahadevan. Imagination machines: A new challenge for artificial intelligence. In *AAAI*, 2018. 3

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 4

[8] Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018. 4, 5, 8

[9] Carl G Jung. Collected works of cg jung: The archetypes and the collective unconscious (vol. ix), 1968. 5

[10] Noam Chomsky et al. *New horizons in the study of language and mind*. Cambridge University Press, 2000. 5

[11] 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. 5

[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. 5, 31

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. 7, 14

[14] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 1989. 8

[15] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1991. 8

[16] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 8

[17] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv*, 2015. 9

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv*, 2017. 9

[19] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017. 11

[20] H. Reichenbach. *The Direction of Time*. University of California Press, 1956. 11

[21] Judea Pearl and Dana Mackenzie. *The Book of Why*. Hachette Book Group, 2018. 12

[22] 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. 12

[23] Tijmen Tieleman. *Optimizing neural networks that generate images*. University of Toronto (Canada), 2014. 13, 14

[24] Thomas G Bever and David Poeppel. Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics*. 14

[25] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*. 14

[26] 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. 14

[27] Guillaume Desjardins, Aaron Courville, and Yoshua Bengio. Disentangling factors of variation via generative entangling. *arXiv*, 2012. 14

[28] 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. 14, 16

[29] 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. 14, 16

[30] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. *ICLR*, 2018. 14

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

[32] Emily L Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017. 14, 16, 31

[33] David A Ross and Richard S Zemel. Learning parts-based representations of data. *JMLR*, 2006. 15

[34] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychol. Rev.*, 1987. 15

[35] Pedro F Felzenszwalb, Ross B Girshick, David A McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 15

[36] David Novotny, Diane Larlus, and Andrea Vedaldi. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *CVPR*, 2017. 15

[37] Saurabh Singh, Abhinav Gupta, and Alexei A Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 15

[38] Grégoire Mesnil, Antoine Bordes, Jason Weston, Gal Chechik, and Yoshua Bengio. Learning semantic representations of objects and their parts. *Mach Learn*, 2013. 15

[39] 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. 15

[40] Michael Lam, Behrooz Mahasseni, and Sinisa Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2017. 15

[41] Tu Dinh Nguyen, Truyen Tran, Dinh Q Phung, and Svetha Venkatesh. Learning parts-based representations with nonnegative restricted boltzmann machine. In *ACML*, 2013. 15

[42] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. In *ECCV*, 1998. 15

[43] Yue Wu and Qiang Ji. Robust facial landmark detection under significant head poses and occlusion. *CVPR*, 2015. 15

[44] 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. 15

[45] Xiang Yu, Feng Zhou, and Manmohan Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 15

[46] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *TPAMI*, 2016. 15

[47] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015. 15

[48] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014. 15

[49] 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. 15

[50] Catalin Ionescu, Fuxin Li, and Cristian Sminchisescu. Latent structured models for human pose estimation. In *ICCV*, 2011. 15

[51] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 15

[52] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015. 15, 25

[53] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 15

[54] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016. 15, 20, 25

[55] 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. 15

[56] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 15, 30, 32

[57] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 15, 19, 22, 23, 25

[58] Karel Lenc and Andrea Vedaldi. Learning covariant feature detectors. In *ECCV Workshops*, 2016. 15

[59] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *NIPS*, 2017. 15

[60] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 15, 19, 23, 25, 26, 29, 30

[61] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 15, 23, 25, 28, 31

[62] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 16, 30, 31, 32, 33, 34

[63] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 16, 31

[64] 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. 16, 31

[65] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017. 16, 31

[66] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. *CVPR*, 2018. 16

[67] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John V Guttag. Synthesizing images of humans in unseen poses. *arXiv*, 2018. 16

[68] Zejian Li, Yongchuan Tang, and Yongxing He. Unsupervised disentangled representation learning with analogical relations. In *IJCAI*, 2018. 16

[69] Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv*, 2018. 16

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 20

[71] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 23

[72] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 23

[73] 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. 25

[74] 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. 25

[75] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 25

[76] 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. 25

[77] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 30, 31

[78] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 30, 31

[79] 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. 31

[80] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*. IEEE, 2017. 31

[81] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 31