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 acts as a mirror, in which these elements of the world are reflected and leave patterns. To recognize these patterns in an image, means 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. Formally this can be posed as an inference problem: a number of latent variables $\mathbf{z_1} \dots \mathbf{z_N}$ has interacted in certain ways to cause the existence of the observed image $\mathbf{x}$. An inference algorithm aims at recovering these latent variables from the observation, *i.e.* the image. These recoveries can be seen as an estimate $\hat{\mathbf{z}}_\mathbf{i}$ for - or a representation of - the true latent variables $\mathbf{z_i}$. A graphical model of the process is shown in figure 1.1. A disentangled representation should then represent each causal element and its state independently: A change in the real causal element $\mathbf{z_i}$ should correspond to an equivalent change in the abstract representational factor $\hat{\mathbf{z}}_\mathbf{i}$, while leaving the other factors $\hat{\mathbf{z}}_\mathbf{j}, j \neq i$, that represent other causes, unchanged.

Typically, objects appear in an intricated interaction of many factors of variation. multiple levels of interaction For example, given the object class of people, variation can be in visual appearance such as the persons clothing or skin color or variation in geometric shape determined by a persons pose or body physique. For articulated object classes the most prominent factors are geometric shape and visual appearance. Disentangling these factors is a difficult problem due to the intricated interplay of shape and appearance under articulation. The complexity enters, as a variation in shape is a change of the images domain rather than a change of its values Shu et al. [2018]. 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 **?**. 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 goal in its own. Once disentangled, a factor can be manipulated individually to make a targeted change. This enables machines to reason about the world Pearl [2018], 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 managable tasks for human imagination, but are out of

Figure 1.1: (a) Causal elements $\mathbf{z_i}$ contributing to generation of image $\mathbf{x}$, (b) Shape $\mathbf{s}$ and appearance $\mathbf{a}$ influencing the image of an object.

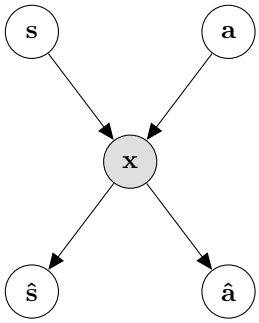


Figure 1.2: Disentangling causal factors: *e.g*. inferring an estimate of object shape $\hat{\mathbf{s}}$ and appearance $\hat{\mathbf{a}}$ from an image

the league for currently used generative image models Goodfellow et al. [2014], Kingma and Welling [2013], that typically rely on uninterpretable vector spaces with entangled dimensions. In the sense of generative modelling, disentangling factors could as well lead the way from a science of images to a science of imagination Mahadevan [2018].

1.2 How not to disentangle?

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 abstract example to illustrate this point: 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, appearance)? Lets suppose the distribution of varation follows a two-dimensional Gaussian distribution, one dimension for pose, one for appearance. The learning algorithm has access to random samples 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 varition in the dataset. But are the two dimensions that the algorithms finds disentangled? No. In fact, a linear combination of pose and appearance and its orthogonal complement are equally valid. Just from observing a two-dimensional Gaussian, no meaning will be attached to the axes. In practice, this problem is often circumvented in-

Figure 1.3: The image captions are generated by a deep neural network (Neuraltalk2) Karpathy and Fei-Fei [2015]. Apart from "common sense" psychology, an understanding in terms of causality is absent Tenenbaum [2018]. Instead, correlating associations are captioned. Human ability to narrate a story needs a causal treatment.

terpolating in the latent space afterwards and determining 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 arbitrary causal factor. The information about how to select the axes, *i.e.* which factors separate, is not contained in raw data. statistical Fitting a model to the data distribution, does in general not yield insights about how the data was generated.

1.3 How can humans disentangle?

Humans have access to richer data, than randomly sampled images from a data set. They observe the world in a temporal sequence, which already reveals a lot about how factors change and persevere across time. Most importantly, humans interact with their environment. And anyone observing a human baby play can affirm that learning humans are obsessed with interaction.

3. Learning by interacting: knowing change by changing. second rung on causal ladder (Pearl): intervention. (, acting) What happens if I do? P(s, do(a)) Others: counterfactual (imagining), association. In humans e.g. egomotion cues: how does image on retina change if I move. *Interaction is crucial.* $\rightarrow$ barometer example: How to find out the causal connection between a barometer and the weather. There cannot be an abstract intelligence, which finds out about the world purely by observation. The intelligence has to interact with the world, it has to be in the world. before this becomes too philosophical infer causation from correlation RCT

1.4 How to disentangle?

change factor → image change equivariantly, leave others invariant → equivariance, invariance

change can be mimicked artificially Intelligent pattern recognition algorithms, fuelled by sensory data as learning material alone, may ultimately drive the way to a full-blown artificial intelligence, reasoning about the world on its own. - That is the reasoning behind data-driven and assumptionless machine learning approaches that have conquered several research communities. A theoretical objection to driving-only-with-data comes from the causal literature: For an understanding of the world, an algorithm needs to model causal processes, that cause an image to be generated.

1.5 Contributions

In this thesis, two hypotheses are proposed and validated: **Hypothesis *i)***: Learning object shape requires abstracting away the objects appearance. This is aided by a disentangled generative modelling of both factors. **Hypothesis *ii)***: Learning disentanglement from raw data without any assumptions is fundamentally constrained Pearl [2018]. In accordance with the causal literature: disentangling causal factors will need assumptions on causal model and/or interactional data (instead of raw data). need to interact with the world, need to change, need to model physical reality -> image transformations, analyis-by-synthesis To test these hypotheses, we *explain*, *validate* and *evaluate* a method for unsupervised shape learning, developed by Lorenz *et al*. 2018.

To explain, we give an overview over state-of-the-art disentangling methods, situate the proposed method in relation to these and analyze future directions and important aspects for disentangling causal factors.

To validate, we show that the proposed method outperforms the state-of-the-art for unsupervised learning of object shape on various datasets, featuring human and animal faces and bodies. We also contribute self-made video datasets for disentangling human pose and appearance, for articulated animal motion and for articulated composite objects (pair dancing salsa). We highlight challenges of the different datasets and how the method tackles them.

To evaluate, we perform a ablation studies on critical components of the method. In particular, we compare to a method which does make the goal of disentangling explicit. In addition, we evaluate the disentanglement performance against a shape-supervised state-of-the-art disentanglement method and perform favorably, proving that a disentanglement is in fact achieved. In short, the results are an above average performance in unsupervised object shape learning. The first disentanglement of articulated object shape from its appearance.

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 observing the data **?**. The source of data points x is usually understood as a probability distribution $x \sim p(x)$.

2.1.1 Supervised

Supervised learning denotes the task to learn a mapping from data points x to target labels y. A supervised algorithm has access to data-label pairs $(y, x) \sim p(y, x)$, to estimate $p(y|x)$, or a function $y = f(x)$. The label can be discrete e.g. a object class label or continuous e.g. the location of an object part in an image.

2.1.2 Unsupervised

why?:? usage of unlabeled data -> find structure in data space; transfer learning, multi-task learning Unsupervised learning is the endeavour to learn structure and patterns in unlabelled data. The learning algorithm then has access to the data distribution $x \sim p(x)$. The task is usually framed as a form of density estimation, to model the entire distribution $\hat{p}(x)$. Thus, after estimation one can generate samples from this model $\hat{p}$, which is then called generative model, cf. sec. 2.2.

model-free vs model-based rigid enough to be useful, flexible enough to useful recently data-driven -> flexible

limits of unsupervised learning? how much prior modelling should be employ? -> as much as possible as long as it is good? (link post Inference)

modeling data distribution P(y, x) sampling from distribution possible e.g. outlier detection P(X) has low probability

2.1.3 Neural Network Models

Artificial neural networks (NN) are a powerful and flexible tool for function approximation. They are inspired by biological neural networks. A function $y = f(x)$ with vector

input $x = \{x_i | i = 1 \dots n\}$ and vector output $y = \{y_j | j = 1 \dots m\}$ is modelled by:

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

, with weight matrices w, w', non-linear 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 are considered *deep* if they comprise multiple hidden layers a la $h_j = a(\sum_i w_{ji} h_i + b_i)$. It can be shown theoretically, that in the limit of infinite hidden units h_j, that NN can approximate any (continuous) function arbitrarily close **?**. In practice, however deeper networks seem to work better. For processing image data, one constrains the weight matrices to be only locally connected and to share weights across locations to enforce translation invariance, resulting in *convolutional* neural networks.

feature hierarchy **?** optimization via gradient descent has proven successful (for deep networks called backpropagation)

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 the data distribution, is the rationale behind generative modelling. Generative models which are mostly applied for unsupervised learning and can be distinguished from discriminative models, that are used to model posterior conditionals $p(y|x)$ **?**. extend on discriminative The currently predominant formulations for generative models are build on autoencoding or adversarial formulations: talk about data compression

2.2.1 Autoencoding

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, is 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 and variance) based on the image.

2.2.2 Adversarial

Generative Adversarial Networks (GAN) 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 discriminator can also be interpreted as a learned similarity metric to measure the closeness of an image to the data distribution. (**?**) The generator is then optimized to make the output indiscriminable from the data distribution.

2.3 Disentangling Causal Factors

In supervised learning a performance measure is naturally given by the metric that is optimized. In the unsupervised setting, judging the performance of a model less straightforward. For example, when modelling an image domain, one could subjectively rate the quality of the generated image. But what characterizes the quality of the latent representation? cite representation quotes The latent representation z learned by generative models captures the essential features of the data distribution.

2.3.1 Equivariance and Invariance

factors should

2.4 Impediments to Causal Learning

The type of knowledge that can be gained by learning from data is limited: so far fitting curve p(x) to data manifold what is missing to human-level intelligence? (cite lake 2016)

causal learning is a hard problem: instead of only learning statistical measures from data, model also needs to be learned (cite Schoelkopf)

Hypothesis: disentangling factors = estimating causal factors -> needs causal for estimation of causal factors "raw data" insufficient -> need interventional data or model assumptions. we do both: 1. intervene with changes to an image which are assumed to change only one factor. 2. model the causal process of the image generation in the theme of analysis-by-synthesis

what does the causality literature have to say? lacking the tools to accurately estimate causality, researchers shied away from making causal statement. Developing machines with human-like abilities requires discovery and reasoning in terms of causal models. Recently (in the past 30 years), overshadowed by the prominent success of data-driven deep learning, the field of causality has emerged to mathematical rigor.

- ladder of causation: association, intervention, counterfactual - current machine learning mostly on level of association (correlations estimated from "pure" data) -> purely data-driven approach can only go so far humans seem to have innate assumptions on coherence, causality, physics etc. introducing inductive biases

measure: p(x) assume causal model: p(x | a, s) want: p(s) and p(a)

encoding $p(s) = p(s|x)\ p(a) = p(a|x) = p(a|s, x)$

decoding $p(x) = p(x|a, s)p(a)p(s)$

p(x| do(s), do(a))

example: Gaussian only with access to p(x) hopes to find factors p(a, b) = p(a) p(b) (InfoGAN, BetaVAE) what if not full-filled? two-dimensional Gaussian: axis x and y are independent factors. in general any superposition of x and y which is orthogonal, can be found imagine a perfect dimensionality reduction yielding a two-dimensional latent space one can find the axes that correlate most with human understanding of independent factors i.e. pose and appearance. But how can a machine find these axes automatically from raw data? it cant, neither can anyone (including humans) (Pearl). Humans know these factors are independent from observing that they can change independently e.g. from observing someone undressing or changing his pose (i.e. harnessing temporal information, with the assumption of temporal coherence) or by changing the factors themselves e.g. what happens to the image of me if I change my pullover? It can be proven mathematically (Pearl) that interventional data or at least certain (which) causal assumptions about the world are necessary to estimate certain quantities.

we harness intervention p(x| do(a), b)

3 Literature Review: Disentangling

research for papers connecting disentangling and causality

3.1 Learning Object Shape

for estimating shape s from images x the task is $p(s|x)$ representation of shape can be landmarks

Disentangling generative factors definition model-free vs model-based approaches:

- model-based $\rightarrow$ more flexible, transferable, modular combination (like parts)

parts as regional attention (cite attention paper) parts/compositionality is key to creativity -> new combination of known parts

3.2 Analysis-by-Synthesis

Capsules, Tieleman make model as good as we can implementing as many assumptions as we can and only leave the rest to powerful model Synthesis known, analysis only indirectly by observing cognition

leaving synthesis to learning from scratch, can meet practical/computational limits *e.g.* convolutional neural networks better than fully connected neural models. But can also be ultimately impossible. Modelling synthesis explicitly with a causal model about image generation, by knowledge about the physical world enables answering interventional and counterfactional questions. (mathematically impossible to learn from "pure" data alone)

3.3 Causal Learning

Pearl Ladder of Causation: rung one seen, rung two seeable, rung three cannot be seen. Barometer example, best neural network will not know -> theoretical proof (look up in Pearl) -> need interaction with the world / or causal assumptions.

3.4 Disentangled Generative Models

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 Goodfellow et al. [2014] and autoencoding Kingma and

Welling [2013] model formulations. Recently, the endeavour for disentangling explanatory factors in the latent representation is being made explicit in the objective functions Burgess et al. [2018], **?** 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. Denton:2017uf

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 Esser et al. [2018], Ma et al. [2017b], de Bem et al. [2018], Ma et al. [2017a], Siarohin et al. [2018], Balakrishnan et al. [2018]. While most of them explain the object holistically, only few also introduce a factorization into parts Siarohin et al. [2018], Balakrishnan et al. [2018]. 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 Higgins et al. [2017], Chen et al. [2016], Li et al. [2018], Denton and Birodkar [2017], Shu et al. [2018], Xing et al. [2018]. 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.

3.6 Part-based Representation Learning

Describing an object as an assembly of parts is a classical paradigm for learning an object representation in computer vision Ross and Zemel [2006] with linkage to human perceptual theories Biederman [1987]. 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 Felzenszwalb et al. [2010], Novotny et al. [2017], Singh et al. [2012], Mesnil et al. [2013], Yang et al. [2016], Lam et al. [2017]. 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 Ross and Zemel [2006], Nguyen et al. [2013], Cootes et al. [1998].

3.7 Landmark Learning

There is an extensive literature on landmarks as compact representations of object structure. Most approaches, however, make use of manual landmark annotations as supervision signal Wu and Ji [2015], Ranjan et al. [2017], Yu et al. [2016], Zhang et al. [2016], Zhu et al. [2015], Zhang et al. [2014], Pedersoli et al. [2014], Ionescu et al. [2011], Toshev and Szegedy [2014], Pfister et al. [2015], Wei et al. [2016], Newell et al. [2016], Lim et al. [2018], Cao et al. [2017].

To tackle the problem without supervision, Thewlis *et al*. Thewlis et al. [2017b] proposed enforcing equivariance of landmark locations under artificial transformations of images. The equivariance idea had been formulated in earlier work Lenc and Vedaldi [2016] and has since been extended to learn a dense object-centric coordinate frame Thewlis et al. [2017a]. However, enforcing only equivariance encourages consistent landmarks at discriminable object locations, but disregards an explanatory coverage of the object.

Zhang *et al*. Zhang et al. [2018] 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*. Jakab et al. [2018] 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 (Thewlis et al. [2017b], Zhang et al. [2018], Jakab et al. [2018]) we take the extend of parts into account, when formulating our equivariance constraint. Furthermore, we explicitly address the goal of disentangling shape and appearance on a part-based level by introducing invariance constraints.

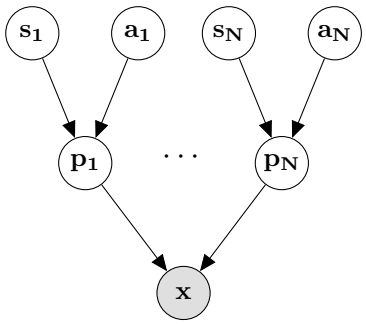
4 Method

> *disentangle as many factors as possible, discarding as little information about the data as is practical* - Y. Bengio, A. Courville and P. Vincent **?**

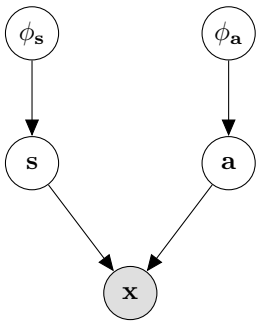
To capture an object in an abstract representation, we follow two key ideas: *(i)* disassembling the object into its constituent parts and *(ii)* disentangling spatial geometry (shape) from visual features (appearance). Hence, we model an object as a composition of parts, each part with a part appearance and a part shape, see Fig. **??**. 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.
local features theme 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.

4.1 Framework

We want to represent an object in an image X. Let us denote the part shape with p_X and the part appearance with f_X. For an object with n parts, the overall shape is constituted by the collection of its part shapes $p_X = (p_X^1, ..., p_X^n)$, the same goes for the appearance $f_X = (f_X^1, ..., f_X^n)$. 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.say the following or not?
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 π, which, if applied to an image X, directly act on the underlying pixel space Λ. Along the same lines we can define appearance transformations ϕ, which act on



(a) Modelling an image of an object as a collection of parts, each with its own shape and appearance



(b) Implementing the do-operation with a transformation of factors

Figure 4.1

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:

$$f_{\pi(X)} = f_X \qquad \text{(invariance of appearance)}$$
$$p_{\phi(X)} = p_X \qquad \text{(invariance of shape)}$$
$$p_{\pi(X)} = \pi(p_X) \qquad \text{(equivariance of shape)}$$

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 X not from the shape and appearance (f_X, p_X) determined from the original image X, but from appropriately transformed images $(f_{\pi(X)}, p_{\phi(X)})$. If the invariance constraints, as formulated above, are fullfilled, these transformations do not change the latent code. Thus, the loss implicitly enforces invariance. To obtain shape and appearance, we encode both $\phi(X)$ and $\pi(X)$ with an encoder E. And, after a recombination R (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:

$$\mathcal{L}_{\text{rec}} = \|X - \mathrm{D}[\mathrm{R}(f_{\pi(X)}, p_{\phi(X)})]\| \tag{4.1}$$

Let us examine what this formulation means on the level of a single part: the part appearance f_X^i is extracted at locations in the spatially transformed image $p_{\pi(X)}^i$, but then used for reconstruction at the location in the original image p_X^i. 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.

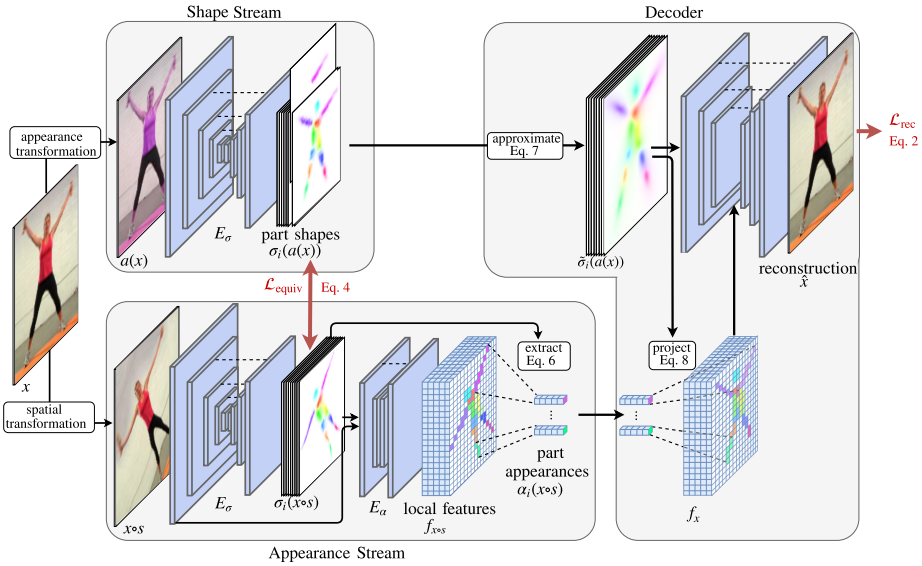


Figure 4.2: Encoder E encodes shape and appearance for two images $\pi(X)$ and $\phi(X)$, after recombination R of $(f_{\pi(X)}, p_{\phi(X)})$ into latent image Z, the decoder D reconstructs the image X.

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 **??** 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 Architecture

The auto-encoding pipeline consists of three stages, namely: **encoding** both shape and appearance for each part, **recombining** this information meaningfully into a latent image and **decoding** this latent image to reconstruct the image. The whole process is sketched in Fig. 4.2, the operations in more detail are visualized in Fig. **??**, **??**, **??**, **??**. 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

$(f, p|X)$[1] The encoding of shape and appearance given an image proceeds in two steps:
(i.) $(p|X)$: The part shapes are predicted given the image. To extract part shapes we use an hourglass[2] neural network: The input is an image X, the output a stack of n part shapes $s = \{p^i | i = 1, ..., n\}$.
(ii.) $(p|f, X)$: The part appearances $f = \{f^i | i = 1, ..., n\}$ 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: $f^i = \sum_{x \in \Lambda} A(x) \frac{p^i(x)}{\sum_{x' \in \Lambda} p^i(x')}$. Each f^i now describes the appearance of a part spatially localized by the part shape p^i.

4.2.2 Recombination of Factors

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 remerge 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: $z^i(x) = p^i(x) \cdot f^i$. Secondly, we reassemble the object from its parts: the part feature images z^i are summarized by summing in a single image: $Z(x) = \sum_i \frac{z^i(x)}{1 + \sum_j z^j(x)}$. The result is an image of part feature descriptors located according to their corresponding part shape, which we call latent image Z.

[1] For a slim notation, we leave out the explicit reference to the generic input image X in this section: f, p, f^i, p^i refer to f_X, p_X, f^i_X, p^i_X.
[2] We utilize hourglass neural network models in both steps, as these models are able to preserve pixel-wise locality, but integrate information from multiple scales **?**.

4.2.3 Synthesis

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 upsampling stream of a standard U-Net . The latent image is scaled to different scales alter figure z1 etc and inserted, after each layer, in addition to the part shapes why add 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 $\hat{p}^i(x) = \frac{1}{1+(x-\mu)^T\Sigma^{-1}(x-\mu)}$, where μ and Σ are the mean and the covariance matrix of the part shape p^i. This allows to pass second-order information such as size and orientation of the part to the decoder. Note that this operation are fully differentiable.

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 the parts exist at multiple scales. For landmark learning, we approximate the part shapes in the decoder in the bottleneck also with $\hat{p}^i(x) = \frac{1}{1+(x-\mu)^T\Sigma^{-1}(x-\mu)}$, but fix the covariance Σ to 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.this needs further explanation or is dangerous
To instantiate shape transformations π, 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 ϕ is encompassing a colour augmentation, contrast variations, and changes in brightness. In general, the more selective the transformation distinguishes shape and appearance, the more invariant the representation.

5 Experiments

5.1 Shape Learning

5.1.1 Landmark Discovery

Human Faces

Human Bodies

Human, Olympic, Penn

Animal Faces/Bodies

Dogs, Cats, Birds

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.

Object/Background Separation

Complexly cluttered background is actually favorable for the method. Correlations of object with background will belong to object.

5.1.2 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

Color, Contrast, Hue

5.1.3 Natural Changes

Video data: Penn, Own

5.2 Disentangling Generative Factors

t-SNE of Shape Representation

5.2.1 Disentangling Pose and Appearance

ReID

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

Pose

PCK Curve

5.2.2 Factorizing into Parts

Own Dataset: Move KP DeepFashion: exchange parts

5.3 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

6 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

7 Bibliography

Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John V Guttag. Synthesizing images of humans in unseen poses. *arXiv preprint arXiv:1804.07739*, 2018.

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

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

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

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.

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

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 preprint arXiv:1804.06364*, 2018.

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

Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018.

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

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.

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.

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

Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013.

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

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

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

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.

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

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2017b.

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

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

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. *ECCV*, 2016.

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

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

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

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.

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

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.

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

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.

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

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

Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018.

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

James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017b.

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

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

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

Xianglei Xing, Ruiqi Gao, Tian Han, Song-Chun Zhu, and Ying Nian Wu. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv preprint arXiv:1806.06298*, 2018.

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.

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

Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018.

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

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

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