Contents

1 Introduction 2

1.1 Why Disentangle Causal Factors? 2

1.2 How not to Disentangle. 3

1.3 How can Humans Disentangle? 4

1.4 Contributions 5

I Appendix 6

A Datasets 7

B Lists 8

B.1 List of Figures 8

B.2 List of Tables 8

C Bibliography 9

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.

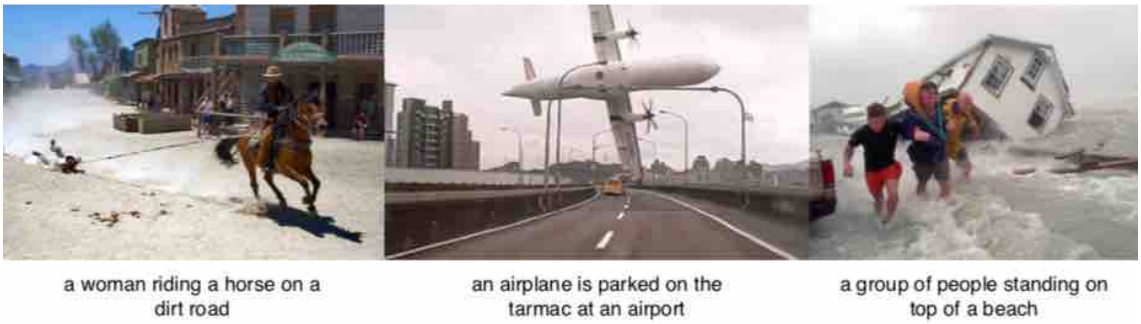
a woman riding a horse on a dirt road
an airplane is parked on the tarmac at an airport
a group of people standing on top of a beach

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 a causal model is absent [8]. Instead, the neural network seems to capture mere associations.

1.2 How not to Disentangle.

Can machines tell a story? 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 shapes complement, 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.

Part I

Appendix

A Datasets

CelebA [13] 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 [14]. As [14, 15], 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 [16] has nearly 9k images of cat heads. We use the train-test split of [15] for training (7,747 images) and testing (1,257 images). We regress 5 of the 7 (same as [15]) annotated landmarks. The images are cropped by bounding boxes constructed around the mean of the ground truth landmark coordinates and resized to 128×128.

CUB-200-2011 [17] comprises ca. 12k images of birds in the wild from 200 bird species. We excluded bird species of seabirds, roughly cropped using the provided landmarks as bounding box information and resized to 128×128. We aligned the parity with the information about the visibility of the eye landmark. For comparing with [15] we used their published code.

BBC Pose [18] contains videos of sign-language signers with varied appearance in front of a changing background. Like [19] 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 [19], we utilized the provided evaluation script, which measures the PCK around $d = 6$ pixels in the original image resolution.

Human3.6M [20] features human activity videos. We adopt the training and evaluation procedure of [15]. For proper comparison to [15] we also removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset.

Penn Action [21] contains 2326 video sequences of 15 different sports categories. For this experiment we use 6 categories (tennis serve, tennis forehand, baseball pitch, baseball swing, jumping jacks, golf swing). We roughly cropped the images around the person, using the provided bounding boxes, then resized to 128×128.

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

Deep Fashion [22, 23] 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 [24]) and used the provided train-test split. For comparison with Esser *et al.* [25] we used their published code.

B Lists

B.1 List of Figures

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 a causal model is absent [8]. Instead, the neural network seems to capture mere associations. 3

B.2 List of Tables

C 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. 2

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

[3] Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *WSDM*, 2018. 2, 5

[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. 2

[5] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *ICLR*, 2013. 2

[6] Sridhar Mahadevan. Imagination machines: A new challenge for artificial intelligence. In *AAAI*, 2018. 2

[7] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 3, 8

[8] Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018. 3, 4, 8

[9] Carl G Jung. Collected works of cg jung: The archetypes and the collective unconscious (vol. ix), 1968. 4

[10] Noam Chomsky et al. *New horizons in the study of language and mind*. Cambridge University Press, 2000. 4

[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. 4

[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. 4

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 7

[14] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 7

[15] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *CVPR*, 2018. 7

[16] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection - how to effectively exploit shape and texture features. In *ECCV*, 2008. 7

[17] 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. 7

[18] 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. 7

[19] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Conditional image generation for learning the structure of visual objects. *NIPS*, 2018. 7

[20] 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. 7

[21] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*, 2013. 7

[22] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 7

[23] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Fashion landmark detection in the wild. In *ECCV*, 2016. 7

[24] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 7

[25] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. *CVPR*, 2018. 7