Contents

1 Prerequisites on Learning Disentanglement

1.1 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 Goodfellow et al. [2016]. 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 $x_i \sim p(x)$.

1.1.1 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). insert example for landmarks Recent advances, in particular the effectiveness of neural network models (cf. sec. 1.1.3) 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, many supervised 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 requiring 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.

1.1.2 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 probabilitstic model (cf. sec. 1.2).

much harder also unspecified finish section connection between unsupervised and supervised learning, cite dlb.

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 where $p(x)$ has low probability

talk about data compression what does unsupervised even mean? no prior assumptions, no knowledge at all? unspecified.. read on this. notion of truly unsupervised learning actually harmful to progress, ill-defined -> intro

1.1.3 Artificial Neural Networks

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

$$\begin{aligned} h_j &= a(\sum_i w_{ji} x_i + b_i) \\ y_j &= a'(\sum_i w'_{ji} h_i + b'_i) \end{aligned} \tag{1.1}$$

, 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 theoretically, that in the limit of infinite hidden units h_j they can approximate any (continuous) function arbitrarily close **?** cite other. In practice, however, networks with more that one layer, referred to as deep neural networks, seem to work better. feature hierarchy citezeiler14vis

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.

longstanding model gained hype-status as working together optimization via gradient descent has proven successful (for deep networks called backpropagation) differentiable

1.2 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. 1.1.1), generative models capture the complete data distribution $p(x)$ in an estimate $\hat{p}(x)$citebishop06ml. Thus, after estimation, one can generate samples from this model $\hat{p}$, hence the name generative model. The currently predominant formulations for learning generative models are built on either autoencoding or adversarial formulations: useful for outlier search

1.2.1 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:

$$\mathcal{L} = \mathbb{E}_{z \sim q(z|x)} \log p(x|z) - \mathrm{KL}(q(z|x)||p(z)) \tag{1.2}$$

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.

1.2.2 Adversarial Formulations

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:

$$\min_D \max_G -\frac{1}{2}\mathbb{E}_{x \sim p_{data}}[\log D(x)] - \frac{1}{2}\mathbb{E}_{z \sim p(z)}[\log(1 - D(G(z)))] \tag{1.3}$$

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.

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 Isola et al. [2017]. add triple gan, ...

1.3 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. For example, when modelling an image domain, one could subjectively rate the quality of the generated image. But even for a qualitative assessment the question arises, how to rate the quality of the latent representation?

1.3.1 Learning Representations

> Disentangle as many factors as possible, discarding as little information about the data as is practical. - Bengio et al. [2013]

According to Bengio et al. [2013] 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 - so goes their reasoning.
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:

1.3.2 Disentangling as Equivariance and Invariance

The definition of factor by change static ... factors should represent elements of real world - change in element -> corresponding change in representational factor - leave other factors representing other elements invariant

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 estimates $\mathbf{\hat{z}_i}$ 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 $\mathbf{\hat{z}_i}$, while leaving the other factors $\mathbf{\hat{z}_j}, j \neq i$, that represent other causes, unchanged.

mathematically,.. $f \circ g(x) = ...$

1.4 Theoretical Impediments from Causality

factors are causal As outlined earlier, the type of knowledge that can be gained by learning from "raw" data is limited. With raw data we mean data x sampled from a $p(x)$. so far

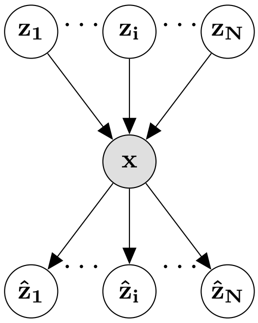

Figure 1.1: Disentangling causal factors means to infer an estimate - *i.e.* a representation - from an image

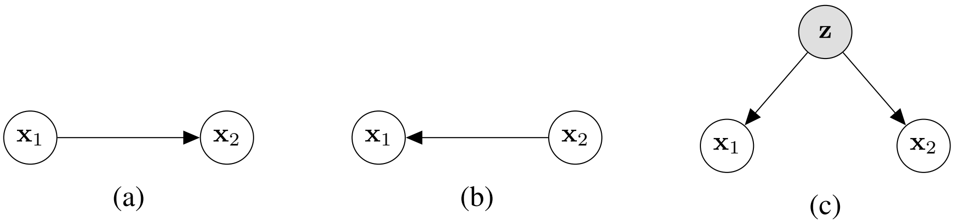

Figure 1.2: Correlation implies causation - if x_1 and x_2 correlate, a) x_1 may cause x_2, b) x_1 may be caused by x_2 or c) both are contingent on a latent cause z

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 (Peters et al. [2017])

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

1.4.1 Causal Learning requires Interventions or Assumptions

What does the causality literature have to say? Statistic background → correlation is not causation. Reichenbachs principle Reichenbach [1956] → barometer example: How to find out the causal connection between a barometer and the weather. Highly capable machine learning algorithm that learns only with access to an image dataset showing the barometer and the weather. -> will be able to capture the correlation between needle position and weather condition, but never understand causal direction, since it is not in 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. - model of influences (humidity -> barometer) What if no prior knowledge? A child-level simple solution is to force the needle to move with a finger. The weather will not change. Hence causality has to go other way or third latent variable influencing both. - intervening: move barometer needle by hand -> no change in weather, hence causality has to go the other way, (example from Pearl and Mackenzie [2018]) 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

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.

1.4.2 Interventions are Transformations

we harness intervention p(x| do(a), b) in computer vision an intervention is an image transformation if ..

1.4.3 Assumptions in Analysis-by-Synthesis

Inverse graphics

1.5 Object Shape and Appearance

2 Literature Review: Disentangling

research for papers connecting disentangling and causality

2.1 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 → 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

2.2 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)

2.3 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.

2.4 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

2.5 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.

2.6 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].

2.7 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.

3 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.

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

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 Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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.

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

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

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.

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 and Dana Mackenzie. *The Book of Why*. Hachette Book Group, 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.

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

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.

H. Reichenbach. *The Direction of Time*. University of California Press, 1956.

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.

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.