Imperial College of Science, Technology and Medicine
Department of Computing

# Unsupervised Learning of Human Movement From Images

Luca Jan Schmidtke

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the Diploma of Imperial College, January 2025

# Abstract

Capturing and understanding human movement through machines has many applications including animation, virtual- and augmented reality, sports as well as medicine. All these benefit from the advent of increasingly powerful, data-driven artificial neural networks with millions or billions of parameters. This advancement, however, is accompanied by an ever-surging demand for more computational resources.

Data with which intelligent systems can be trained comes in two forms: with or without annotations. In the supervised training paradigm, models can be trained to predict these annotations coming from human input. Data without annotations however is many times more abundant. Humans have the remarkable ability to learn an internal world-model, *i.e.*, a rich set of rules that govern the world in which they live in from different sensory inputs without such supervision. A central goal of artificial intelligence research is to enable machines to do something similar.

The central topic of this thesis is to investigate *how to capture human movement from images through self-supervision.* This movement can be characterised by *pose*, a set of points that denote the location of anatomical landmarks such as skeletal joints. The central theme throughout all the chapters is learning a disentangled representation of *pose* and *appearance* of any person in an image though a computational model.

In this thesis we find that a human shape template, a collection of connected, basic shapes, that serves as an abstract representation for pose, provides a strong prior for unsupervised pose estimation. Different self-supervision mechanisms such as conditional image translation as well as generation are introduced in combination with this shape template. We find that we can learn human pose and appearance from images in a flexible way and with minimal supervision, paving the way towards machine learning systems that can leverage unprecedented amounts of data.

## Acknowledgements

I would like to acknowledge the following wonderful people without whom none of this would have been possible.

First of all, I want to thank my supervisor Bernhard Kainz for his unwavering support and commitment in our discussions and meetings, in improving my manuscripts and providing the perfect conditions for not just me, but many other PhD students as well.

A lot of thanks goes to my wife Melat, who not just supported, but also encouraged me to develop myself in ways I would not have achieved on my own.

I want to thank my parents Ute & Lothar for letting me grow up in a home that I can describe with no other words than comfort and warmth, a home that I want to create for my children in the exact same way.

I also thank the rest of my family and friends for their companionship throughout my life. Especially Jonas for being a good brother that sometimes provides advice on the practical side of life.

Lastly, I also thank my former colleagues at Imperial, especially Thanos and Benjamin, for fruitful discussions and their contributions to my work.

# Dedication

To my daughter Hannah Marie (and any future children of ours): *The future of this world is yours to shape.*

# Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Statement of originality

I hereby declare that all contents of this thesis are my own work and any work from others has been properly acknowledged.

Luca Schmidtke, March 2024

# Contents

**Abstract** **i**

**Acknowledgements** **iii**

**1 Introduction** **1**

1.1 Motivation and objectives . . . . . 1

1.2 Contributions . . . . . 3

1.3 Publications . . . . . 5

**2 Background and literature review** **6**

2.1 Normalization layers in neural networks . . . . . 6

2.1.1 Batch normalisation . . . . . 6

2.1.2 Instance normalisation . . . . . 7

2.1.3 Adaptive instance normalisation (AdaIN) . . . . . 7

2.1.4 Spatially adaptive denormalisation (SPADE) . . . . . 8

2.2 Generative models . . . . . 8

2.2.1 Generative adversarial networks . . . . . 10

2.3 Image-based human motion analysis . . . . . 14

2.4 Image-based human pose estimation . . . . . 15

2.4.1 Fully supervised pose estimation . . . . . 15

2.4.2 Weakly- and unsupervised human pose estimation . . . . . 16

2.4.3 3D pose estimation . . . . . 18

2.5 Pose estimation in medicine . . . . . 19

2.5.1 Quantifying infant movement . . . . . 20

2.6 Modelling the human body . . . . . 22

2.6.1 2D body-models . . . . . 22

2.6.2 3D-body-models . . . . . 23

2.7 Summary . . . . . 26

**3 2D-Pose Estimation Via Differentiable Shape Templates 27**

3.1 Introduction . . . . . 27

3.2 Contextual overview . . . . . 28

3.3 Data collection . . . . . 30

3.4 Method . . . . . 30

3.4.1 Template and anchor-points . . . . . 31

3.4.2 Pose extractor . . . . . 33

3.4.3 Image translation module . . . . . 33

3.4.4 Training objectives . . . . . 33

3.5 Experiments . . . . . 35

3.5.1 Evaluation procedure . . . . . 36
3.5.2 Results . . . . . 37
3.6 Downstream application study . . . . . 40
3.7 Summary . . . . . 41
3.8 Appendix . . . . . 42
3.8.1 Network architectures . . . . . 42
3.8.2 Additional qualitative results . . . . . 43
3.9 Training and evaluation details . . . . . 44
3.9.1 Human3.6m . . . . . 44
3.9.2 Infants . . . . . 44

**4 3D-Pose Estimation via Neural Rendering 45**

4.1 Introduction . . . . . 45
4.2 Contextual overview . . . . . 46
4.3 Method . . . . . 47
4.3.1 Template and volume rendering . . . . . 48
4.3.2 Pose regression and shape transformation . . . . . 49
4.3.3 Kinematic chain . . . . . 50
4.3.4 Loss function . . . . . 51
4.3.5 Infant-adjusted SMPL model . . . . . 51
4.4 Experiments . . . . . 52
4.5 Results . . . . . 53

4.6 Summary . . . . . 56

**5 Pose Estimation Via Generative Modelling 57**

5.1 Introduction . . . . . 57

5.2 Contextual overview . . . . . 58

5.3 Related work . . . . . 60

5.4 Overview . . . . . 61

5.4.1 Generative pathway . . . . . 62

5.4.2 Training objectives . . . . . 68

5.5 Experiments . . . . . 70

5.5.1 Datasets . . . . . 71

5.5.2 Evaluation . . . . . 72

5.5.3 Results . . . . . 72

5.5.4 Ablation Studies . . . . . 73

5.6 Additional experiments . . . . . 75

5.6.1 Random shifts . . . . . 75

5.6.2 Disentanglement of pose and appearance . . . . . 77

5.6.3 Estimating pose through inversion . . . . . 77

5.6.4 Frozen generator . . . . . 79

5.6.5 Improving segmentation . . . . . 83

5.6.6 Results on infants . . . . . 88

5.7 Summary . . . . . 89

5.8 Appendix . . . . . 89

5.8.1 Training and hyperparameters . . . . . 89

5.8.2 Architecture . . . . . 90

**6 Discussion & Conclusion** **93**

6.1 Discussion . . . . . 93

6.2 Outlook . . . . . 96

6.3 Conclusion . . . . . 98

**7 Appendix** **99**

7.1 Computer vision, machine learning & deep learning . . . . . 99

7.1.1 Artificial neural networks . . . . . 101

7.1.2 Convolutional neural networks: A strong prior for images . . . . . 102

7.2 Optimisation . . . . . 104

7.2.1 Gradient descent . . . . . 104

7.2.2 Backpropagation and automatic differentiation . . . . . 106

7.2.3 Other common components of neural networks . . . . . 107

7.2.4 Residual networks . . . . . 108

**Bibliography** **109**

# List of Tables

3.1 Chapter 3: a table comparing access to labels of related works . . . . . 38

3.2 Chapter 3: Quantitative results on Human 3.6M . . . . . 39

3.3 Quantitative results on infant dataset . . . . . 39

3.4 Chapter 3: ablation study . . . . . 40

3.5 Chapter 3: 3d lifting results . . . . . 41

5.1 Chapter 5: quantitative pose estimation results on Human3.6m . . . . . 74

# List of Figures


2.1 Spatially adaptive instance normalisation . . . . . 8

2.2 Normalisation schemes . . . . . 9

2.3 Generative adversarial networks: training method and plot of cost functions . . 11

2.4 StyleGAN architecture . . . . . 12

2.5 GAN inversion . . . . . 13

2.6 Human pose estimation example outputs . . . . . 19

2.7 General movement assessment: example longitudinal study . . . . . 21

2.8 Pictorial structures model . . . . . 23

2.9 SMPL: rendered meshes for different shapes and body configurations . . . . . 25

3.1 infant dataset collection setup . . . . . 30

3.2 Chapter 3: Method overview . . . . . 31

3.3 Chapter 3: architecture overview . . . . . 32

3.4 Chapter 3: results on Human3.6M . . . . . 38

3.5 Chapter 3: failure cases . . . . . 39

3.6 Chapter 3: qualitative results . . . . . 40

3.7 Chapter 3: 3d pose lifting . . . . . 41

3.8 Appendix chapter 3: architecture of pose extractor . . . . . 43

3.9 Appendix chapter 3: architecture of image decoder . . . . . 43

3.10 Appendix chapter 3: additional pose estimation results on Human3.6m . . . . . 43

3.11 Appendix chapter 3: qualitative pose estimation results on infant dataset . . . . 44

4.1 Chapter 4: method overview . . . . . 47

4.2 Chapter 4: illustration of the kinematic chain . . . . . 50

4.3 Chapter 4: SMPL model adjusted to infants . . . . . 52

4.4 Chapter 4: soft rasterisation . . . . . 53

4.5 Chapter 4: qualitative results on Human3.6m . . . . . 55

4.6 Chapter 4: mesh-based SMPL model results . . . . . 55

4.7 Chapter 4: qualitative results on infant dataset . . . . . 56

5.1 Chapter 5: overview figure . . . . . 61

5.2 Chapter 5: network architecture overview . . . . . 62

5.3 Chapter 5: shape template . . . . . 63

5.4 Chapter 5: reconstruction task and U-net . . . . . 67

5.5 Chapter 5: qualitative results on Human3.6m . . . . . 71

5.6 Chapter 5: occlusion-aware part maps . . . . . 73

5.7 Chapter 5: synthetic pose-conditioning . . . . . 73

5.8 Chapter 5: qualitative results on YouTube dataset . . . . . 74

5.9 Chapter 5: ablation study . . . 75

5.10 Chapter 5: ablation study . . . 76

5.11 Chapter 5: disentanglement of pose and appearance . . . 78

5.12 Chapter 5: inversion results with frozen generator . . . 79

5.13 Chapter 5: PsP encoder . . . 80

5.14 Chapter 5: foreground and background inversion . . . 81

5.15 Chapter 5: additional inversion objectives . . . 81

5.16 Chapter 5: joint inversion . . . 82

5.17 Chapter 5: inversion loss diagram . . . 83

5.18 Chapter 5: Qualitative results of mask refinement model trained on Human3.6m. 85

5.19 Chapter 5: minimising spillage between fore- and background . . . 87

5.20 Chapter 5: loss curves during training . . . 87

5.21 Chapter 5: qualitative results on the infant dataset . . . 88

5.22 Appendix chapter 5: Detailed architecture of the foreground generator . . . 91

5.23 Appendix chapter 5: Additional generative results of the model trained on Human3.6m . . . 91

5.24 Appendix chapter 5: additional generative results of the model trained on the YouTube dataset. . . . 92

7.1 Schematic display of artificial neuron . . . 100

7.2 Image convolution visualised . . . 104

7.3 Schematic overview of the AlexNet architecture . . . 108

# Chapter 1

# Introduction

## 1.1 Motivation and objectives

Tracking and understanding human movement from images is a long-standing problem in computer vision. Applications range from motion tracking for animating digital characters in film or video games to virtual- and augmented reality, optimisation of biomechanics in sports or gait analysis in the medical context. Thanks to advances in both scalable algorithms in the form of artificial neural networks and accompanying hardware, image-based automated tracking of human movement has become more and more accurate and reliable, especially in the two-dimensional case. At the same time, efficient neural network architectures and optimised hardware can nowadays even power real-time applications on handheld devices such as smartphones. This in turn enables enormous potential in different areas including healthcare, where the automation of labour-intensive tasks such as image analysis can aid health practitioners and therefore drive down costs and increase accessibility of diagnostic tools across the population. However, many models require large amounts of human-labelled images in order to train and fall short in terms of performance when there is domain shift between training and actual data distribution encountered during deployment. The latter often brings about a need for fine-tuning models on additional data that lies closer to the specified domain. Self-supervised learning methods on the other hand are aimed towards finding underlying patterns in data

without relying on human supervision in the form of labels. This in turn enables tapping into much bigger sources of data: for computer vision, the number of easily accessible images on the internet is virtually endless. Additionally, as demonstrated in the recent leaps in performance in language processing as well as computer vision applications Devlin et al. [2019], Caron et al. [2021], pre-training a large model with a lot of parameters in a self-supervised fashion with unprecedented amounts of data can dramatically increase performance in subsequent applications when being fine-tuned for specific tasks afterwards. These so called foundation models have become ubiquitous in both language processing and computer vision, especially in the context of generative modelling.

In the context of human pose estimation, previous work has addressed the question whether pose landmarks can be inferred without direct supervision, but rely on various forms of weaker forms of supervision such as an independent, unpaired set of annotations Jakab et al. [2020a] or require a mapping of discovered landmarks to anatomically relevant landmarks via a linear classifier.Lorenz et al. [2019a], Zhang et al. [2018a]. In this thesis we explore the boundaries of what can be achieved in a self-supervised setting by pushing the amount of required supervision, whether direct or indirect, to an absolute minimum.

Beyond the practicality of not relying on manual labels, self-supervised learning constitutes a key-component of intelligence. Intuitively, interacting with and forecasting events within the world we observe on a daily basis requires that we build an internal model that explains it. In a more technical sense, a model that aims at matching or replicating an observed probability distribution over data can reveal underlying structure. In the context of this thesis, the underlying structure we would like to discover in images is the *decompositionality of the human anatomy into connected parts and the location of the points that connect them.*

The central research question of this thesis therefore is: Can we extract said underlying structure from images accurately in order to minimise the extend of supervision in the context of human pose estimation?

In more detail, the following criteria should be met:

- no reliance on manually defined pose landmarks during training and testing
- sufficient inference speed and manageable model size
- highest possible accuracy, flexibility and temporal consistency of detected landmarks

The body of research presented in this thesis follows this research question at its core, but can be split into conceptually different approaches and sub-goals that will be laid out in detail in individual chapters.

## 1.2 Contributions

### Unsupervised human pose estimation through transforming shape templates

Firstly, in the second chapter we explore the usage of a simple two-dimensional body model consisting of connected and transformable geometric primitives in combination with a self-supervision mechanism to learn human pose representations from images. We demonstrate on-par performance on a common human pose estimation dataset with other methods relying on various forms of additional weak supervision[1] We also show that our method remains flexible when being trained on infants instead of adults, where different body proportions and pose distributions lead to a significant domain shift that reduces the performance of comparable methods. However, the applicability to more diverse scenarios remains constrained by the need for constant backgrounds during training.

[1]Unsupervised, self-supervised and weakly supervised learning are labels being used somewhat laxly by many authors. See 2.4.2 in the background chapter.

### Extension to 3D via neural rendering

In the third chapter we introduce an extension of the previous chapter to a more challenging problem: estimating pose in three dimensions solely based on two-dimensional images. To tackle this we leverage the same self-supervision mechanism in combination with a three-dimensional volumetric shape template and a specific rendering technique referred to as neural rendering. We show with qualitative results that this is a promising future direction of research, but also encounter difficulties due to depth-ambiguity that call for a modified approach with additional weak supervision.

### Combining shape template transformation with a generative approach for unconstrained self-supervision.

In the fourth chapter we address the central limitation of the self-supervision framework introduced in the previous two chapters by enabling training on different backgrounds. An analysis-by-synthesis approach is introduced, where a generative model is trained to synthesise images via a composition of disentangled factors including pose without any additional supervision. This is achieved through a combination of novel objective functions within the context of Generative Adversarial Networks and a new depth-aware representation to tackle self-occlusion in two-dimensional body-models. We also highlight some limitations of the method including imperfect foreground-background separation as well as difficulties from training on synthetic data and provide extensive experiments that point towards possible future research.

## 1.3 Publications

### Included in this thesis

**L. Schmidtke**, A. Vlontzos, S. Ellershaw, T. Arichi, A. Lukens, and B. Kainz, "Unsupervised Human Pose Estimation through Transforming Shape Templates", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021*

**L. Schmidtke**, B. Hou, A.Vlontzos, and B.Kainz, "Self-Supervised 3D Human Pose Estimation in Static Video Via Neural Rendering", *CV4Metaverse Workshop at European Conference on Computer Vision (ECCV) 2022*

### Not included in this thesis

Yueqi Cao, Anthea Monod, Athanasios Vlontzos, **Luca Schmidtke**, and Bernhard Kainz, Topological information retrieval with dilation-invariant bottleneck comparative measures, *Information and Inference: A Journal of the IMA, Volume 12, Issue 3, September 2023*

Simon Ellershaw, **Luca Schmidtke**, Nidal Khatib, Jonathan Eden, Sofia Dall'Orso, Silvia Muceli, Etienne Burdet, Niamh Nowlan, Tomoki Arichi, and Bernhard Kainz, *Medical Imaging Meets NeurIPS Workshop at Conference on Neural Information Processing Systems, 2020*

### Unpublished

**L. Schmidtke**, B.Kainz, "Towards Estimating Pose Via Analysis-By-Synthesis", submitted to CVPR 2024 but rejected

# Chapter 2

# Background and literature review

In order to embed the presented work into a fitting context, the following sections serve as a short introduction and summary of the relevant literature connected to our research question in this thesis.

## 2.1 Normalization layers in neural networks

### 2.1.1 Batch normalisation

As mentioned in section 7.2.1, the weights of a neural network are updated with subsamples, also called batches, of the training data. Normalising the output of each layer by subtracting the mean and dividing by the standard deviation computed across the batch empirically improves convergence and generalisation performance. Different works [Bjorck et al., 2018, Santurkar et al., 2018, Kohler et al., 2018] have shown that batch normalisation has a smoothing effect on both gradient distributions and loss-landscapes. For a batch of $n$ images with $c$ feature channels, the normalised output is:

$$\tilde{\mathbf{x}}_{n,c} = \gamma_c \frac{\mathbf{x}_{n,c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_c, \tag{2.1}$$

where $\mu_c$ and $\sigma_c$ are the mean and standard deviation computed over all spatial dimensions and across samples of $\mathbf{x}_{n,c}$:

$$\mu_c = \frac{1}{NHW} \sum_{N,H,W} x_{nchw}, \tag{2.2}$$

where $\gamma_c$ and $\beta_c$ are additional learnable scalars and $N$, $H$ and $W$ correspond to number of samples in the batch and height and width respectively.

### 2.1.2 Instance normalisation

Instead of normalising across all samples in a batch, in instance normalisation the mean and standard deviation are computed for each sample individually:

$$\mu_{nc} = \frac{1}{HW} \sum_{H,W} x_{nchw} \tag{2.3}$$

Its main application lies in style transfer [Johnson et al., 2016, Gatys et al., 2016, Huang and Belongie, 2017], where global image characteristics (the "style") such as colour-distribution and textural elements are transferred from a target to a source image, preserving the content of the latter. In this scenario, normalising over the whole batch would fail to preserve these image-level statistics.

### 2.1.3 Adaptive instance normalisation (AdaIN)

Instead of learning the affine parameters $\gamma$ and $\beta$ as parameters of the network, adaptive instance normalisation (AdaIN) introduced by [Huang and Belongie, 2017] provides a mechanism to insert statistics of an arbitrary input into the network layers via mapping it to $\gamma$ and $\beta$. Given two sources of input $\mathbf{x}$ and $\mathbf{y}$, where $\mathbf{x}$ is fed through a neural network, an arbitrary feature map $\mathbf{f}^i$ in the $i$-th layer of the network can be conditioned on external input provided

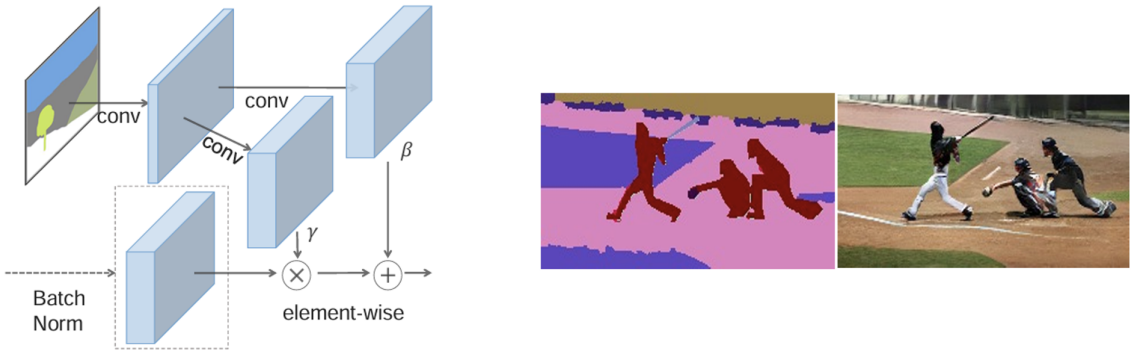



Figure 2.1: Excerpt from [Park et al., 2019b]. Left: architecture of SPADE: spatial input is mapped to modulation feature maps $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ via convolutional layers and ReLU-activations. Right: generated image from semantic map fed into the network layers.

by $\mathbf{y}$ (*e.g.* an image with different style statistics):

$$\tilde{\mathbf{f}}^i_{n,c} = \gamma_{n,c}(\mathbf{y}) \frac{\mathbf{x}_{n,c} - \mu_{n,c}}{\sqrt{\sigma^2_{n,c} + \epsilon}} + \beta_{n,c}(\mathbf{y}), \tag{2.4}$$

### 2.1.4 Spatially adaptive denormalisation (SPADE)

The affine parameters defined in the previous paragraph on AdaIN are scalars and depend on external input $\mathbf{y}$, providing a way to condition network outputs on $\mathbf{y}$. Instead of only using scalars, [Park et al., 2019b] instead introduce $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ as spatially varying tensors that are element-wise multiplied with the feature map. This allows to effectively condition on spatially varying input such as segmentation maps, where granularity on the pixel-level is necessary. Figure 2.1 illustrates the concept.

## 2.2 Generative models

In its most general form, a generative model aims to approximate a joint distribution $p(\mathbf{x}, \mathbf{y})$ of data $\mathbf{x}$ and associated labels $\mathbf{y}$ of any shape and form. There is a wide range of direct applications including content generation in form of text, images and video. However, learning how to generate samples from an observed distribution over data can aid in learning structure

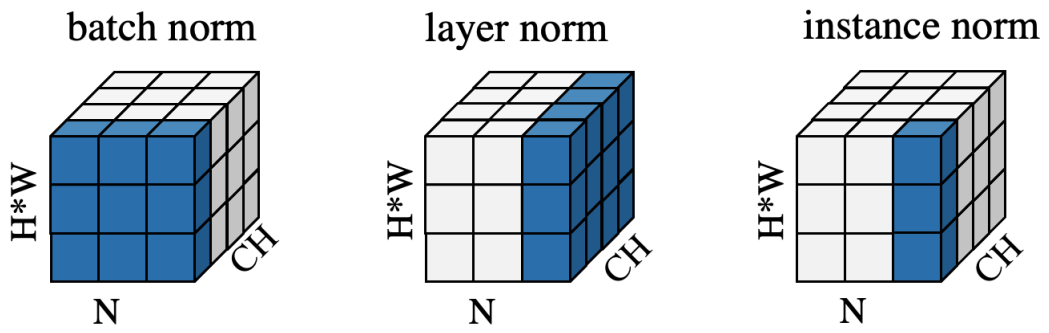



Figure 2.2: Different normalisation schemes based on a typical tensor with shape (N, CH, H*W), where N is the batch size, CH denotes the channels and H*W are height and width from an unrolled two-dimensional tensor.

within it without relying on external supervision. Generative modelling can be a form of unsupervised learning tapping into a huge reservoir of freely accessible unlabelled digitised data through the internet. While there are many different approaches, it is helpful to structure generative models into implicit and explicit density estimation as done in [Goodfellow, 2016]. Assuming a target data distribution $p(\mathbf{x})$, explicit models define a parametric density $p_\theta(\mathbf{x})$ with parameters that can be directly optimised towards matching the target distribution, most commonly via the *maximum-likelihood* approach. The idea is to maximise the likelihood of the training data under the model distribution with respect to parameters $\theta$:

$$\theta^* = \arg\max_\theta \frac{1}{n} \sum_{i=0}^{n} \log(p_\theta(\mathbf{x}_i)) \tag{2.5}$$

For complicated models with many parameters such as neural networks and high-dimensional data such as images, computing that likelihood directly quickly becomes intractable. An alternative approach are implicit generative models, where a (typically low-dimensional) stochastic variable $\mathbf{z} \in \mathbb{R}^m$ from a prior distribution $q$ is mapped to a target distribution via a deterministic function $f_\theta : \mathbb{R}^m \to \mathbb{R}^n$ (*e.g.* a neural network):

$$\mathbf{x}_g = f_\theta(\mathbf{z}'),\ \mathbf{z}' \sim q(\mathbf{z}) \tag{2.6}$$

The space in which the stochastic variable is defined is called a *latent space* and represents a condensed, lower-dimensional representation of the typically high-dimensional data that the

generative model is tasked to learn. If neural networks serve as a mapping between latent and data space and back, this process is often hidden from our direct view and remains a black box. Despite this however, recent works have shown surprising amounts of interpretability in the structure of latent spaces of neural networks in the form of simple vector arithmetics [Radford et al., 2015] and linear interpolations [Upchurch et al., 2016].

### 2.2.1 Generative adversarial networks

A particular class of neural-network based implicit generative models has achieved big leaps in the generation quality of image and video data: Generative Adversarial Networks (GANs) [Goodfellow et al., 2014a]. The general idea is the following: A so called generator network $G_\theta$ maps a low-dimensional stochastic variable from some prior distribution (usually a standard normal distribution) $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{z} \sim \mathcal{N}(0, 1)$ to a sample $\mathbf{x} \in \mathbb{R}^n$ of the generated distribution $p_g(\mathbf{x})$. A second network $D : \mathbb{R}^n \to \mathbb{R}$, coined the discriminator, maps either a sample from the distribution $p_r(\mathbf{x})$ of real, observed data or the generated distribution to a scalar and is trained to classify samples based on whether they come from either of the distributions. The generator on the other hand acts as an "adversary" in the optimisation sense and is trained to maximise the cost function of the discriminator. In other words, the discriminator is trained to distinguish real samples from generated ones, while the generator tries to "fool" the discriminator into classifying generated samples as real with random chance. This can be formalised via a minimax game of the following cost funtion:

$$\max_\theta \min_\phi \mathcal{L} = -\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log(D(\mathbf{x}))] - \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log(1 - D(G_\theta(\mathbf{z}))] \tag{2.7}$$

#### Modified objectives

As argued in [Goodfellow, 2016], the cost function for the generator can produce gradients approaching zero (and therefore stopping meaningful convergence via gradient-descent) when the discriminator classifies generated samples as fake with high confidence. One alternative is

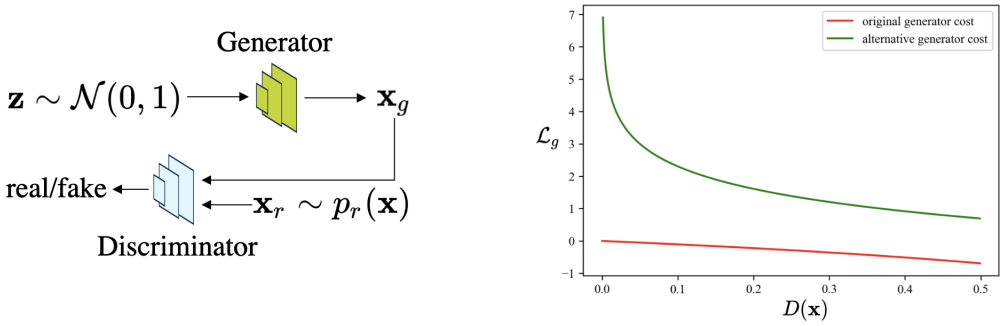



Figure 2.3: Left: schematic overview of adversarial training including the generator and discriminator networks. Right: Cost of the generator as a function of the discriminator output. The original objective exhibits a very flat curve around 0, that is, when the discriminator rejects generated samples with high confidence.

to flip the targets in the cross-entropy loss instead of the sign of the cost function:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log(D(G_\theta(z))] \tag{2.8}$$

It has been demonstrated empirically that this alternative objective improves convergence [Goodfellow et al., 2014a]. Despite the improved quality of generated images, especially with regards to sharpness compared to other models such as variational autoencoders [Kingma and Welling, 2013] or autoregressive generation [van den Oord et al., 2016], GANs are plagued by a different set of problems:

- **Mode collapse**: The assumed probability distribution over natural images is multimodal, *i.e.* there are several regions where a lot of probability mass (relative to the support of the distribution) is concentrated around. Adversarial training often leads to a phenomenon coined *mode collapse* in which the generator defaults to only produce samples from one of these modes. In practice this leads to a lack of variety of generated samples compared to the empirical distribution. In severe cases, it might even lead to the generator mapping all samples of the latent variable to the same image.

- **Unstable training dynamics**: GANs are notoriously difficult to train depending on which distribution should be learned. Oftentimes the discriminator overpowers the generator and achieves near perfect classification very early in training despite the modified

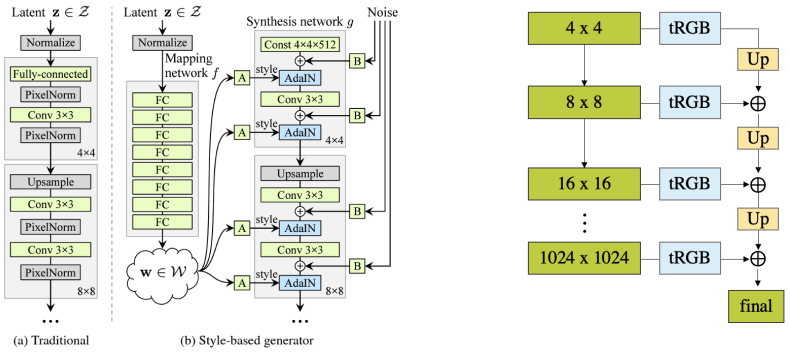



Figure 2.4: Left: Standard and StyleGAN architecture Karras et al. [2018]. The latent variable is mapped to a "style" vector that is fed into every layer. Right: modification to the StyleGAN architecture. Progressive growing is replaced by skip connections. "Up" denotes an upsampling operation to match resolutions between outputs.

objective or training does not converge to a meaningful output.

There is a rich set of literature on addressing both of these issues in GANs [Arjovsky et al., 2017, Nguyen et al., 2017, Mao et al., 2016, Metz et al., 2016, Miyato et al., 2018], but neither can fully alleviate them. In fact, it has been observed in many experiments that sample quality and diversity are inversely related and require a trade-off [Karras et al., 2018].

### StyleGAN

Alongside improvements on the cost function and different normalisation schemes, one particular modification of the architecture of the generator network has led to big improvements in generated image quality. StyleGAN [Karras et al., 2018] borrows the mechanism of Adaptive Instance Normalisation (AdaIN) (see 2.1.3) to inject the latent code into each layer and enabling fine-grained control over the generation process at different scales. Different from [Goodfellow et al., 2014a], the upscaling pathway starts with a learnt constant and the learning process is stabilised by progressively growing the spatial resolution of the output during training. Another iteration of this architecture [Karras et al., 2019b] replaces the memory-intensive progressive growing of the resolution with an efficient approach based on skip-connections as outlined and motivated in 7.2.4. The final output is a sum of upsampled images extracted from all intermediate resolutions of feature maps throughout the network.

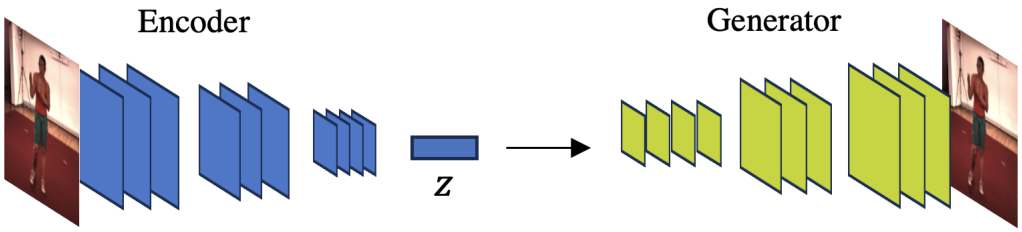



Figure 2.5: Principle of inversion: an input image is mapped into the latent space such that the code results in the same input image when passed through a trained generator network.

### Inversion

Inversion, *i.e.* the reverse mapping from image data to the latent space of the generator network, is an important sub-discipline with many applications in image editing, novel-view synthesis or segmentation. The general principle is displayed in Figure 2.5. After training a generative model, an inversion attempts to map an image to a vector in the latent space of the model, such that it produces the given image in a forward pass through the generator. Given that the structure of the latent space allows for semantic editing, the vector can be changed and mapped back via a forward pass through the generator , leading to local or global changes of certain aspects of the input image. The StyleGAN-architecture in particular has sparked many works attempting inversion and aiming to achieve both faithful image reconstruction and semantic editability [Abdal et al., 2019, Richardson et al., 2020, Alaluf et al., 2021]. Methods can be roughly categorised into two streams, direct iterative optimisation of the latent variable and encoder-based approaches. While the former yields accurate reconstruction results, it is more time-consuming due to its iterative nature. Encoder-based methods only require a single pass through an encoder that maps from image to latent space and are therefore very fast and efficient, but the faithfulness of the reconstructed image to the original is generally lower. Still, due to their fast inference encoders remain the preferable option and we will continue the discussion in that direction.

The StyleGAN architecture (see Figure 2.4) offers several different options for an attempted inversion mapping: the latent space $\mathcal{Z}$ before the mapping through the MLP, the so called *style-space* $\mathcal{W}$ where the style-vector is fed into each layer via adaptive instance normalisation and an extended style-space commonly referred to as $\mathcal{W}^+$, in which the single style vector $\mathbf{w}$ is

extended to a set of style vectors $\{\mathbf{w}_1, ..., \mathbf{w}_n\}$ for each of the $n$ layers in the generator network. In some methods, the weights of the generator network stay fixed during training of the encoder network, in others they are additionally fine-tuned.

## 2.3 Image-based human motion analysis

Human motion analysis in general encompasses the localisation and tracking over time of specific points or sections that are part of the human body. There is a wide range of applications including motion capturing for animated content, kinematic analysis of sports-related movements both at the recreational and professional level, medical diagnosis, surveillance as well general understanding of human behaviour [Aggarwal and Cai, 1999]. The amount of methods and tools for the measuring process is not limited to imagery, but includes wearable optic and magnetic markers of varying degrees of sophistication and depend on the specific use-case and the resources available. Marker-based approaches remain the most accurate and reliable, but come with greatly increased cost and very limited scalability in terms of data acquisition. They require a complex procedure including fitting wearable sensors, sometimes even wired, and connecting them to the measurement unit. This greatly increases the overall cost and resources involved and, combined with the limited amount and the overall weight and dimensions of gear, leads to a restriction to a local setting.

In contrast, image-based tracking only requires a person and a consumer-grade camera, which nowadays can be found in the pocket of nearly every person around the globe as a standard component of a smartphone. Due to recent performance leaps in computer-vision and overall computation power even in hand-held devices, more and more challenges previously deemed impossible to tackle without wearable sensors due to challenging poses, occlusion or non-standardised measuring environments have become or are in the process of becoming feasible candidates for markerless, image-based approaches.

## 2.4 Image-based human pose estimation

Human pose estimation refers to landmark detection applied to human anatomy and is the main tool for motion analysis. Once landmarks have been localised with sufficient accuracy, many of the tasks and applications described in the previous section follow. In general, image-based human pose estimation aims to find a set of points $\{\mathbf{p}_i\}$ in two- or three-dimensional space given an input image $\mathbf{x}$ that correspond to anatomical landmarks of interest. Most commonly, these coincide with joints in the human skeleton including wrists, ankles, shoulders and hips, but can change depending on the definition. Common locations where datasets differ from one another are the exact locations of landmarks around hands and feet.
In this thesis we restrict the problem of predicting the 2D pose of a human from a *single* 2D RGB image as a pose recognition task.

### 2.4.1 Fully supervised pose estimation

Fully supervised models rely on carefully annotated data, which is available for narrow fields of applications, *e.g.* MS COCO keypoints [Lin et al., 2014], MPII Human Pose database [Andriluka et al., 2014], Human3.6M [Ionescu et al., 2014a] or LSP [He et al., 2016b]. Methods utilising these datasets are usually trained without additional priors due to the abundance of direct labels. CNNs have shown evidence to be powerful estimators for keypoints [Toshev and Szegedy, 2014] and their uncertainties [Tompson et al., 2014]. Confidence heatmaps are popular for scenes where a single pose needs to be estimated [Belagiannis and Zisserman, 2017, Bulat and Tzimiropoulos, 2016, Carreira et al., 2016, Newell et al., 2016b, Pfister et al., 2015, Tompson et al., 2015, Wei et al., 2016] or multiple poses at once [Cao et al., 2017, Insafutdinov et al., 2016]. Our framework does not use existing image annotation to learn a pose prediction model.

### 2.4.2 Weakly- and unsupervised human pose estimation

#### A note on terminology

**Unsupervised and self-supervised learning** Throughout this thesis and in fact in all the related literature, the terms *unsupervised* and *self-supervised* learning are often used interchangeably. We will make use of the broadest and most general definition for unsupervised learning, which is learning or extracting patterns from unlabelled (that is without labels provided by some form of human input) data. Within that definition, self-supervised learning can be viewed as a form of unsupervised learning in which labels come from the inherent structure of the input data itself, and most often come with an associated prediction task and a cost function to minimise. This training signal then serves as a proxy for learning powerful representations for later fine-tuning on downstream applications or directly performing a related prediction task that the model has learned to perform from the first task. While the term *unsupervised* learning is clearly more general, the differences to *self-supervision* remain blurry. For example, most experts would categorise autoencoders as an unsupervised learning approach for dimensionality reduction, but it also fits quite well into our definition of self-supervision: the input data itself serves as a label, while the reconstruction from a representational bottleneck is a proxy task for learning condensed representations. For this reason we make no such distinction between the terminologies in the subsequent literature review.

#### Weakly supervised learning

As discussed earlier in this chapter, machine-learning approaches can be divided into subcategories defined by the amount and type of labels they require in order to be trained. Weakly supervised and unsupervised methods including self-supervised approaches have grown in popularity due to their efficiency in dealing with limited ground truth data. This becomes especially important when there are no large public datasets available for the target domain.
We restrict our literature review to weak supervision or no supervision, when only partial or no annotations are available or when models are transferred to a new domain. In the context of

human pose estimation we count methods as either weakly or unsupervised if they make no direct use of manual pose landmarks during the training phase. The question when a method can be referred to as weakly supervised or unsupervised remains debatable, especially since many algorithms rely on certain preprocessing steps such as image cropping according to a bounding box containing the human, which is why again we do not make this distinction explicit.

[Zhang et al., 2018b] introduce an convolutional autoencoder architecture and represent pose landmarks via the grid coordinates of the maxima of spatial heat maps produced by the neural network. They introduce a number of constraints including locality, separation between peaks as well as equivariance under image transformations. Lorenz et al. [2019b] develop the concept of autoencoding further by disentangling object appearance and shape through additional invariance constraints and introduce more complex image transformations, including thin plate-spline image warping.
A key problem in both these approaches is the fact that discovered landmarks do not follow common conventions inspired by human anatomy, *i.e.* ankles and joints, and vary profoundly by method. To tackle this, both make use of a linear classifier that is trained with manual pose landmarks in order to map the model output to said convention. They additionally make use of segmentation masks available in some of the datasets to remove background from the image during training and inference.

[Jakab et al., 2018] propose to learn a pose representation via conditional image translation. By leveraging the extracted pose representation and given another frame containing the same person in a different pose they task the model to reconstruct the original image. However, a pixel-wise reconstruction loss does not encourage meaningful, low-dimensional representations. To avoid the model from simply encoding pixel information similar to an autoencoder, a representational bottleneck in the form of $k$ tuples of $(x, y)$ coordinates transformed to 2D Gaussian heat maps is introduced. This enables unsupervised pose landmark detection provided that the background remains static. However, the extracted landmarks follow no common convention and are difficult or impossible to interpret. To tackle this, [Jakab et al., 2020b] further propose the introduction of a fully differentiable skeleton representation following common human pose landmarks via a distance field from line segments between points and an exponential

fall off. However, adversarial training with unpaired pose labels is required as a form of weak supervision to drive the model towards anatomically plausible pose representations.

[Kundu et al., 2020a] proposed to use an energy-based optimization approach combined with a part-based shape template to estimate 3D poses in images with varying backgrounds. The method produces impressive results but still relies on a set of unpaired real 3D pose labels.

Others propose methods to match pairs of images of an object, but sacrifice geometric invariants such as keypoints to achieve this [Kanazawa et al., 2016, Rocco et al., 2017, Shu et al., 2018]. Sparse and dense landmarks are introduced by [Sundermeyer et al., 2018, Thewlis et al., 2017b, 2019] in the unsupervised context. Synthetic views of 3D models like in [Sundermeyer et al., 2018, Hesse et al., 2015] can be additional sources for weak supervision.

In summary, while all of the above methods either *reduce* or entirely remove the need for image and manual landmark pairs during training, all of them make use of weaker supervision signals such as a linear mapping to anatomically meaningful pose convention during inference, unpaired pose labels from different datasets or by using off the shelf human segmentation models.

### 2.4.3 3D pose estimation

**Monocular 3D pose estimation** Inferring 3D pose from monocular images remains a challenging problem because it is ill-posed. Generally, there is no unique solution for finding points in 3D that project onto given 2D landmarks. But the solution space can be reduced by making use of additional cues in the image or video such as motion or multiple views from synchronised cameras or introducing prior knowledge about the hierarchical part based structure of the human body.

**Lifting from 2D to 3D** Many works break down the problem into first estimating 2D pose and subsequently estimate 3D pose either directly [Martinez et al., 2017a], by leveraging self-supervision through transformation and reprojection [Li et al., 2020] or a kd-tree to find corresponding pairs of detected 2D pose and stored 3D pose [Iqbal et al., 2018].

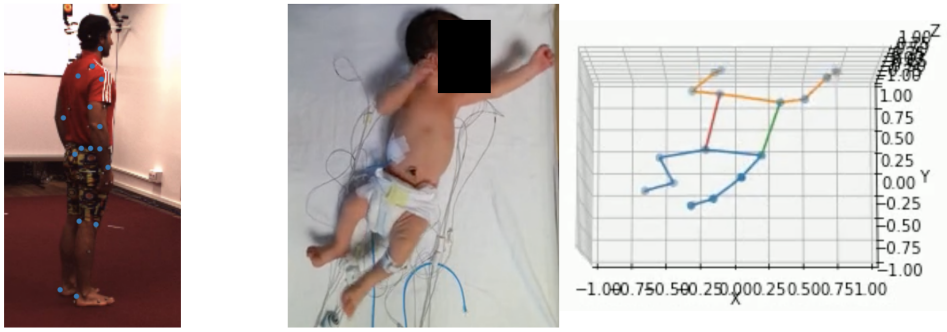



Figure 2.6: Left: estimation of 2D pose on an image from the Human3.6m dataset [Ionescu et al., 2014b]. Right: 3D-pose estimation from a picture of a moving infant from a dataset collected in cooperation with clinical partners. Note that the infant is simultaneously filmed and also connected to a magnetic tracking system via markers placed on the body-locations of interest.

**Motion cues from video** Videos provide a rich source of additional temporal information that can be exploited to limit the solution space. [Lin et al., 2017], [Katircioglu et al., 2018], [Hossain and Little, 2018] and [Kocabas et al., 2020] use recurrent architectures in the form of LSTMs or GRUs to incorporate temporal context while [Pavllo et al., 2019] employ temporal convolutions and a reprojection objective.

**Multiple views** Other approaches incorporate images from multiple, synchronised cameras to alleviate the ill-posedness of the problem. [Pavlakos et al., 2017], [Tomè et al., 2018] and [Qiu et al., 2019] fuse multiple 2D heatmaps while [Rhodin et al., 2018a] and[Rhodin et al., 2018b] utilize multi-view consistency as a form of additional supervision in the objective function.

## 2.5 Pose estimation in medicine

Pose estimation provides a tool to capture and analyse the kinematics of human movement, which in turn can be a valuable bio-marker for the diagnosis, management and prevention of diseases or injuries affecting the human motion apparatus [Stenum et al., 2021]. Examples include abnormal gait detection for the management and detection of stroke [Ferraris et al., 2021] or dementia [Kondragunta and Hirtz, 2020], prevention of work-related injuries [Han and Lee, 2013, Mehrizi et al., 2018] or early assessment of neural development in newborns [Prechtl,

2001]. However, ground truth generation for training supervised machine learning models is laborious and often limited to a narrow domain, for example, standard poses of healthy adults. In the medical field with limited demand or special requirements, extensive labelling efforts are often not justified. Moreover, key point definitions may vary according to diagnostic aims and body shapes might not comply with the learned expectations from a standard training set. Indeed, motion tracking has a variety of applications in medicine, for example to examine the progression of neurological disease and to evaluate treatment success [Kontschieder et al., 2014], the assessment of injury [Bae et al., 2009] or for the early diagnosis of impaired neurological development in infants [Teitelbaum et al., 1998b, Ferrari et al., 2002b]. None of these applications allow for excessive data collection and annotation, often because of a limited number of subjects, restrictions on recording in the clinical environment and economical considerations.

### 2.5.1 Quantifying infant movement

While young infants move spontaneously and without conscious control, the overall pattern or quality of these random movements evolves significantly over the course of the first few months after birth and is a sign for normal and healthy neurological development. A deviation from this trajectory on the other hand is believed to be indicator for adverse conditions affecting neurological and sensorimotor function later in life [Hadders-Algra, 2014]. Anencephaly [Hadders-Algra, 2014] and Cerebral Palsy (CP) [Ferrari et al., 2002a, Einspieler et al., 2012] have been linked to abnormal movement, while other studies suggest links to Autism [Teitelbaum et al., 1998a, Doi et al., 2023] as well. Existing methods for assessing infant movement such as the General Movements Assessment (GMA) [Prechtl, 1990] are very time consuming and resource-intensive, and therefore are not part of the clinical routine, delaying the diagnosis of conditions such as CP significantly. A very early diagnosis and intervention however, are believed to significantly improve the affected children's quality of life due to the greater plasticity of young brains [Novak et al., 2017].

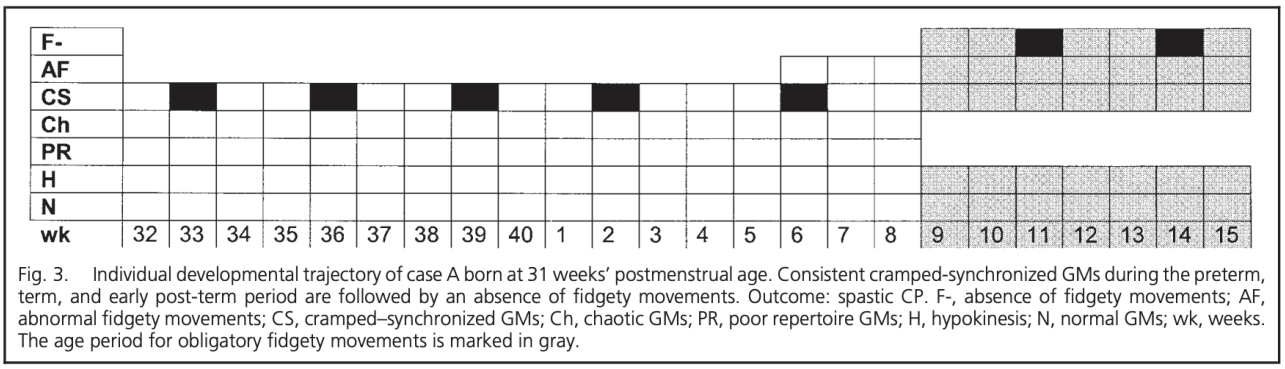



Fig. 3. Individual developmental trajectory of case A born at 31 weeks' postmenstrual age. Consistent cramped-synchronized GMs during the preterm, term, and early post-term period are followed by an absence of fidgety movements. Outcome: spastic CP. F-, absence of fidgety movements; AF, abnormal fidgety movements; CS, cramped–synchronized GMs; Ch, chaotic GMs; PR, poor repertoire GMs; H, hypokinesis; N, normal GMs; wk, weeks. The age period for obligatory fidgety movements is marked in gray.

Figure 2.7: Excerpt from [Einspieler and Prechtl, 2005] displaying the assessment outcomes of several videos of a longitudinal study.

### GMA

GMA involves filming moving infants for a period between 30 to 60 minutes, selecting roughly three sequences of movements making up between 5-10 minutes and classifying them into different categories. This procedure is ideally repeated 4-7 times throughout the first months of the newoborn's life in order to judge the developmental trajectory and overall risk for CP. Figure 2.7 displays an example measurement trajectory of GMA on an assessed infant. While guidelines and workshops have been put into place to standardise the testing environment and conditions, the selecting and rating of appropriate sequences remains somewhat subjective and depends largely on the experience of the judging expert. Additional, more fine-grained score-based testing such as the general movement optimality score (GMOS) [Einspieler et al., 2016] have been suggested in order to better measure the subtleties beyond broad movement categories, which might be necessary to assess the impact of intervention more accurately. Computer-aided methods offer great potential to automate many components in this diagnostic pipeline including data selection in the form of efficient video summarisation or supporting experts in their rating with quantitative and detailed movement tracking and analysis.

### Quantitative, computer-aided methods

In previous work, marker-based approaches relying on optical [Airaksinen et al., 2020] or electromagnetic [Karch et al., 2008] tracking have been utilised to track motion in infants. However,

these methods rely on clinical specialists, expertise, often costly equipment and a substantial amount of manual preparation and calibration. Different marker-less approaches have been proposed based on optical flow [Stahl et al., 2012] or particle matching [Rahmati et al., 2015]. With the advent of compact and cheap cameras with integrated depth sensors, several more recent works combine images and depth information with random fern classifiers [Hesse et al., 2015], a deformable parts model [Khan et al., 2018] or by employing a parametric shape model [Olsen et al., 2014, Hesse et al., 2018b]. More recently, methods based on pose tracking and body-part parsing with CNNs have been proposed [Abbasi et al., 2022, McCay et al., 2020].

## 2.6 Modelling the human body

### 2.6.1 2D body-models

**Geometric primitives** There is a rich set of literature on body models in computer vision, as human anatomy presents a natural prior that can be exploited to solve human pose estimation and segmentation tasks. Early works [Felzenszwalb and Huttenlocher, 2004, Andriluka et al., 2009a, Zuffi et al., 2012] make use of the pictorial structures model [Fischler and Elschlager, 1973], where arbitrary objects including humans are modelled as a collection of deformable geometric primitives such as rectangles and are connected pair-wise. A matching function depending on hand-crafted features and some measure of deformation between connected parts is minimised. [Jakab et al., 2020a] make use of a distance-field function to map from 2D-point coordinates to a skeleton representation. [Kundu et al., 2020a] manually define 2D-parts and connections that can be transformed via affine mappings to match a given pose in an image.

**Contours** Active shape models [Cootes et al., 1995, Freifeld et al., 2010] are another popular approach where a low-dimensional representation of points populating an extracted contour of a number of training examples is found via Principal Component Analysis. Variations in shape are modelled via the first $n$ modes of the principle components, and unseen image contours are fitted into the model-subspace via least-squares.

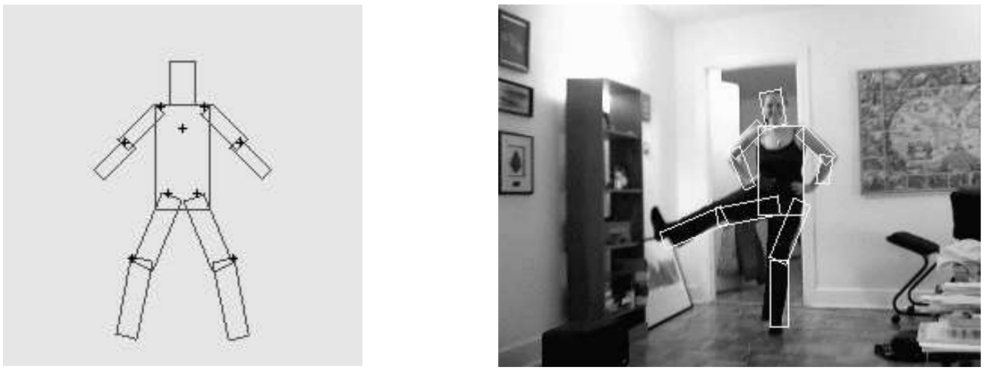


Figure 2.8: Excerpt from [Fischler and Elschlager, 1973]. Left: Pictorial Structures Model with connected parts, right: body model fitted to image.

### 2.6.2 3D-body-models

When it comes to modelling the human body in 3D, the field of computer vision intersects with computer graphics, a discipline concerned with synthesising digital images based on modelling objects and their interactions with light in 3D. In contrast to digital images, where information is stored and modelled via pixels, there is a wide range of different ways to model content in 3D, which is also reflected in the number of different representations used in computer vision applications.

**Mesh-based models** widely used in computer graphics have been adapted for computer vision. The object is represented by a mesh, which is a collection of *vertices* (coordinates denoting a position in 3D as well as holding other information including colour, normal vector and texture coordinates) that are connected by *edges* and are organised into closed sets called *faces*, which, in the simplest and most widely-used scenario, represent triangles. In the rendering process, the mesh is mapped back to two-dimensional image-pixels by a process called rasterisation. Models such as SCAPE [Anguelov et al., 2005] and SMPL [Loper et al., 2015a] make use of triangle meshes in combination with linear skinning, a procedure to tie a group of vertices of a surface to a specific body joint defined in an underlying skeleton structure via a linearly weighted influence of neighbouring bones (connections between joints). Furthermore, identity- and pose-dependent surface deformations are learned separately from one another via Principal Component Analysis from a reference dataset of varying body shapes and poses. Compared to

pictorial structures and silhouette matching, these models are a lot more expressive and are in principle capable of mimicking very detailed, non-rigid deformations of a person's skin in 3D dependent on pose. Many works such as [Kanazawa et al., 2018], [Kolotouros et al., 2019] and [Kocabas et al., 2020] make use of SMPL. [Xu et al., 2020] leverage kinematic constraints to improve their predictions while [Kundu et al., 2020b] leverage a forward kinematics formulation in combination with the transformation of a 2D part-based template to formulate self-supervision in form of image reconstruction similar in some ways to our approach.

**Dense body-part-maps** Given a segmentation or silhouette of a human figure in an image, the human body can be modelled in a more accurate fashion via dense correspondence, *i.e.* assigning each pixel within the segmentation boundary to a predefined body part. [Shotton et al., 2011] demonstrate that this division into separate body parts can improve 3D pose estimation with random forests on depth images when being introduced as an intermediate step. This seminal work was a key component of the Kinect sensor developed by Microsoft, a small consumer-grade device equipped with RGB as well as depth cameras capable of real-time gesture recognition and 3D human pose estimation. While the Kinect has originally been introduced as a motion controller system for a video game console, it has also been used extensively in computer vision research.

[Güler et al., 2018] predict the correspondence between pixels on an image of a human and the 3D surface of a SMPL body [Loper et al., 2015b] by regressing $UV$ coordinates via a convolutional neural network. in a fully-supervised and end-to-end setting. The ground truth coordinates are obtained via a elaborate annotation pipeline involving human annotators and a number of predefined points on the surface. The authors also make use of several intermediate steps including a first coarse classification of each pixel into one of 24 body parts followed by a finer localization within that body part via the $UV$ coordinates. Similar to [**?**], introducing an intermediate body part model as part of a divide-and-conquer-type strategy improves the performance of downstream tasks including pose estimation.

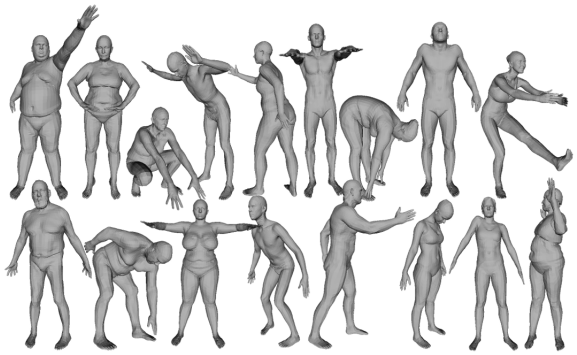


Figure 2.9: Excerpt from [Loper et al., 2015b]. Different rendered body configurations with different shapes and in different poses.

**Voxel-based models** Another common explicit 3D-representation comes in the form of voxels [Jackson et al., 2018, Varol et al., 2018, Zheng et al., 2019], where, analogous to the pixel representation of images, 3D-space is partitioned into a regular grid of cells (voxels), and an object is defined by assigning these voxels and defining their properties. The granularity or resolution of the representation is determined by the voxel size. While straightforward for neural networks to work with, the growth in required memory storage is cubic and quickly becomes impractical for a larger number of voxels.

**Implicit surface models** defined by signed distance functions [Park et al., 2019a, Chen and Zhang, 2018], or occupancy fields [Mescheder et al., 2018b] parameterised by neural networks in combination with fully-differentiable rendering techniques have enabled end-to-end learning of 3D-representations of objects from a collection of 2D images.

**Neural Radiance Fields** [Mildenhall et al., 2020], where an individual object or scene is represented by a radiance field parameterised by a neural network, have further improved performance on 3D object reconstruction.

Both implicit surface models and neural radiance fields have sparked a number of recent works that utilise these representations in 3D human body reconstruction [Mildenhall et al., 2020, Kwon et al., 2021, Su et al., 2021, Jiang et al., 2022, Wang et al., 2022].

## 2.7 Summary

In this chapter we have introduced various concepts including normalisation layers for neural networks, generative adversarial networks, weakly, self- and unsupervised learning as well as human pose estimation and various approaches for modelling the human body. The following chapters will focus on the scientific contributions of this thesis, where the aim is to extract human pose from images with the least amount of weak supervision as possible via a very simple model of the human body combined with two different self-supervision tasks: conditional image translation as well as generative modelling.

# Chapter 3

# 2D-Pose Estimation Via Differentiable Shape Templates

## 3.1 Introduction

As outlined in the overall introduction, the technical focus of this thesis lies in the following question: Can we extract and learn the *decompositionality of the human anatomy into connected parts* from images without relying on supervision through manual labels? If we are able to gain knowledge about the location of these body parts and their relationship with each other from an image, downstream applications such as pose estimation and tracking as well as body-parsing and their subsequent applications in engineering and technology become possible naturally. As briefly touched upon in 2.4, the location of pose landmarks in an image are flexible and depend on the dataset and application. For point-based supervised methods, this means having to re-train the model on a new set of landmarks if the definition changes with the application. One big advantage of a body model is that changing it to alternative landmark positions requires no additional training effort, as, depending on the fidelity of the model, it captures many possible locations on the surface of the human body by design. For example, there may be a need to identify either the wrist or the palm of the hand, necessitating retraining or increasing the amount of labels and therefore the overall annotation cost and effort.

As described in 7.1, self-supervised learning is a popular strategy that relies on finding and training to predict labels that are inherently part of the data itself. One way to learn pose is to model it as a *factor of variation* in different images of the same person combined the task of image translation or reconstruction. The following content of this chapter is based on [Luca Schmidtke and Kainz, 2021] and exploits a simple body-model based on connected geometric primitives combined with self-supervision to find human-interpretable 2D-pose landmarks from single-camera images *without* relying on manual labels for training.

## 3.2 Contextual overview

As outlined in 2.3, human pose estimation is motivated by a wide range of applications including motion capture, visual surveillance and robot control, a continuous effort has been put into generating datasets and models where a manually annotated ground truth pose and key points are available as labels for full supervision. These are currently the most attractive approaches for industrial applications due to their promise of higher accuracy.

As described in the literature review in the previous chapter, unsupervised and weakly supervised pose estimation methods can reduce the manual labelling effort, which is an important factor in the medical domain where annotations are expensive and hard to come by due to privacy concerns. One prominent modelling approach is to decompose images into *appearance*, which encodes individual differences such as clothing or body height and *pose*, describing the individual's positioning and configuration of limbs and joints with a canonical latent code [Jakab et al., 2018, Lorenz et al., 2019b, Kundu et al., 2020a]. In this context, self-supervision tasks, such as image reconstruction or translation, have been shown to be powerful tools to estimate pose as a *factor of variation* across images instead of relying on strong, manual supervision signals.

We therefore aim at learning the 2D geometry of object categories such as humans and infants with no additional supervision. We exploit the structured information provided by raw videos of continuous pose changes and propose to control inductive bias directly for arbitrary

object categories through the manual definition of very simple templates. Thus, we intend to automatically learn a neural representation that, once trained, can predict the 2D pose from a single input image. We show that if such a 2D pose prediction is accurate and compliant with an expected shape prior, these estimations can be extrapolated to 3D poses with a lower error than other existing methods.

We present a method for the unsupervised estimation of 2D keypoints requiring only a simple template and an unannotated video of a single human performing actions in front of a static background to learn a meaningful pose representation. Inspired by previous work [Jakab et al., 2020b], where this problem is framed as a *conditional image generation and translation* approach, pose information is utilised to recover a particular frame of a video from any other randomly chosen time-point. Despite the effective use of self-supervision and representational bottleneck, this approach still requires another prior in the form of unpaired labels and introduces susceptibility to domain shift if these labels come from a different dataset. Our model however does not require an additional dataset of unpaired 2D pose examples and relies solely on a simple 2D template consisting of connected body parts modelled as 2D Gaussians. The update of these Gaussians can be learned as affine transformations. Even though images are 2D representations of 3D information, affine transformations allow to model all possible projected configurations of body parts. Motivated by part-based approaches such as [Zuffi and Black, 2015] we introduce anchor-points in order to enforce connectivity between body parts and regularise model training and prediction.

In summary, we make the following contributions:

- We introduce a conceptually simple but effective method to learn 2D human-interpretable keypoints based on transforming a single manually defined 2D template.
- Our proposed approach is capable of performing 2D human pose estimation without any additional need for labeled data, either paired or unpaired.
- We demonstrate the high adaptability of our approach by evaluating it on benchmark data and in the wild on a challenging infant pose estimation dataset.



Figure 3.1: Left: Cot in the hospital with mounted setup of an Azure Kinect RGB-D camera connected to a tablet for the storage and monitoring. Right: colour image with labelled landmarks in green and corresponding depth-channel image.

## 3.3 Data collection

As part of this project, data was collected in form of RGB-D videos of moving infants in the postnatal ward of St. Thomas' Hospital in London. All data was acquired and shared with ethics approval in relation with EPSRC grant EP/S013687/1. All subjects were recruited from consenting parents. Figure 3.1 displays the setup in the hospital. An Azure Kinect RGB-D camera is mounted onto a hospital cot via a flexible camera mount.

**Labelling** was performed with Labelbox [lab, 2021], an online fully-configurable labelling-tool. In total, three annotators marked the spatial position of the following landmarks (2 for each side): *hips, shoulders, elbows, knees, ankles and wrists.* In total, around 2000 images from 24 subjects were labelled in this way. The labels are subsequently used for training our fully supervised upper bound and for testing all methods displayed in table 3.3.

## 3.4 Method

In this section we present our proposed method. We extend the concept from from [Jakab et al., 2020b] and formulate a self-supervision task by tasking our model to reconstruct an initial frame $\mathbf{f}_t$ conditioned on an estimated low-dimensional pose representation $\tilde{p}_t$. On a high

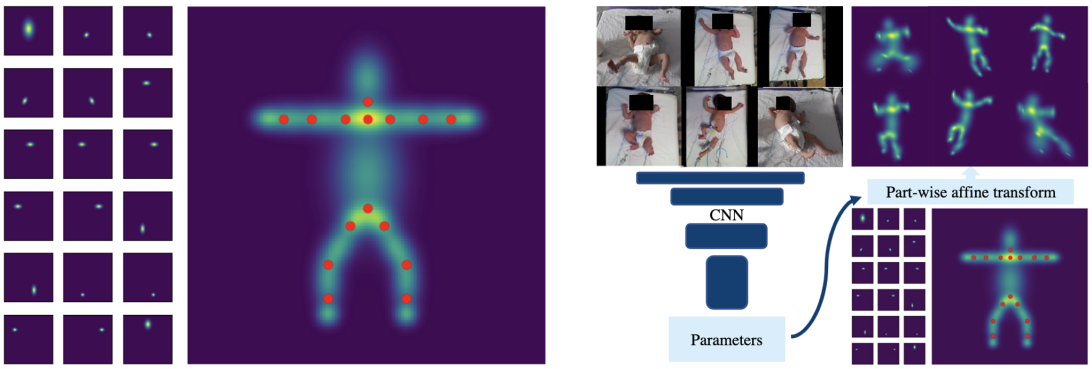



Figure 3.2: Method overview. Left: Template consisting of connected blobs defined by a Gaussian function. Red dots denote anchor points. Right: Conceptual overview of the method. We define a part-based human puppet template and predict the transformation parameters to estimate the pose for an input frame
.

level, our method is composed of two modules. The first network $\varphi$, given the initial frame $\mathbf{f}_t$ and a template $\mathbf{T}$ consisting of 2D Gaussian heatmaps of human body parts, extracts an estimate $\tilde{p}_t$ of the true pose $p_t$ represented by a spatial arrangement of different body parts. The second module is an encoder-decoder network $\phi$ which receives as input the previously extracted pose together with a frame $\mathbf{f}_{t+k}$ containing the same person in a different pose $p_{t+k}$, and tasked to reconstruct the original frame $f_t$.

$$\tilde{\mathbf{f}}_t = \phi(\mathbf{f}_{t+k}, \tilde{p}_t) \tag{3.1}$$

$$\tilde{p}_t = \varphi(\mathbf{f}_t, \mathbf{T}) \tag{3.2}$$

The pose representation $\tilde{p}_t$ is found by optimising the model to encode all necessary information to recover the original pose whilst enforcing representational constraints to avoid the encoding of low-level image features.

### 3.4.1 Template and anchor-points

Inspired by [Kundu et al., 2020a] and [Zuffi and Black, 2015], we design a human template as shown in Figure 3.3 that consists of 18 body parts in an effort to represent human anatomy (forearm, torso, head, etc.). Instead of transforming this template according to previously found

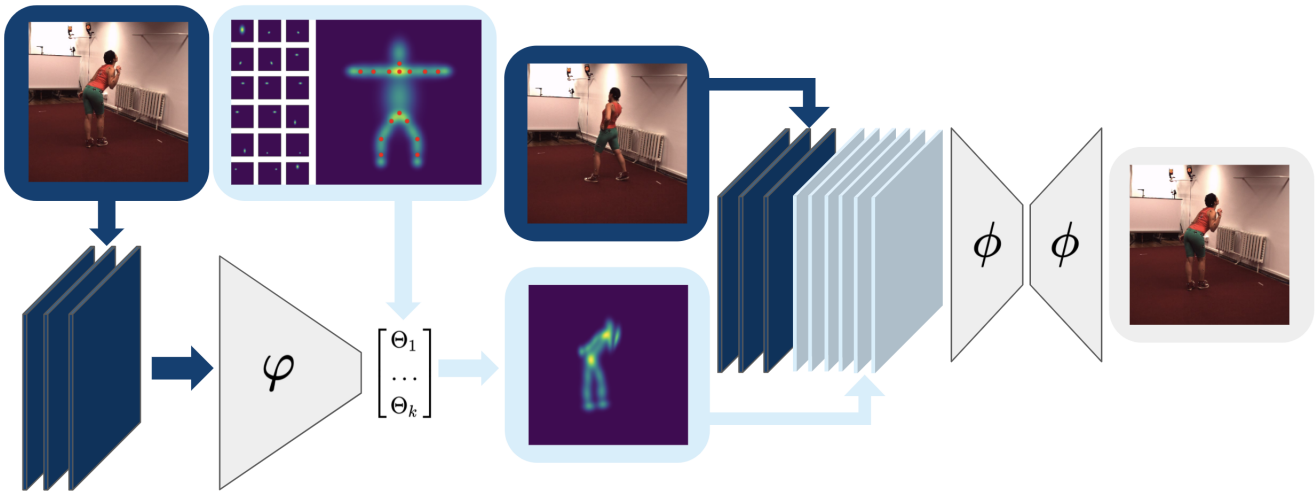



Figure 3.3: The network $\varphi$ regresses transformation parameters for the pose template, the decoder $\phi$ reconstructs this image from another video frame and the transformed template, which it can only achieve if pose information is propagated via the template.

keypoints in the image, we directly estimate the transformations and make use of the template as a strong prior. Each part is characterised by its central position $(x_0, y_0)$ surrounded by a 2D Gaussian with variance $(\text{Var}_x, \text{Var}_y)$. We pre-define the mean and variance of each Gaussian for all body parts in an effort to represent the length and width of the individual parts. Hence, we are able to model a canonical T-pose as shown in Figure 5.3. This pose was chosen because it roughly represents the mean of possible pose configurations. Each representation of a body part $\mathbf{l} \in \mathbb{R}^{h \times w}$ is saved in an individual channel resulting in a tensor of shape $(B, K, H, W)$ where $B$ is the batch size, $K$ the number of body parts and $H, W$ are width and height of the input image.

For each body part $\mathbf{l}_j$ in our template $\mathbf{T}$ we define $N_j$ (up to three depending on which part) anchor points $\mathbf{a}_i^j \in \mathbb{R}^2$ in image space, with $j = 1, ..., K$ and $i = 1, ..., N_j$. Our template design and the resulting anchor points coincide with the most commonly used landmark definitions for human pose estimation. However, the detected landmarks can be easily changed by simply choosing different points on each body part.

### 3.4.2 Pose extractor

The network $\varphi : \mathbb{R}^{3\times h\times w} \rightarrow \mathbb{R}^{k\times 3\times 3}$ is implemented as a fully-convolutional neural network followed by a fully-connected layer. An image $\mathbf{f}_t$ is passed through a series of down-sampling convolutional layers and mapped to a pose representation described by the parameters of affine transformations $\Theta_k \in \mathbb{R}^{3x3}$ for each $k$-th body part from the original template:

$$\Theta = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} & t_x \\ \theta_{2,1} & \theta_{2,2} & t_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.3}$$

where $t_{x,y}$ represent the translations across the two dimensions and $\theta$ correspond to rotation, shear, and scale. Each part of the template is then warped by its corresponding transform, resulting in $\mathbf{T}' = (\vartheta_0(\mathbf{l}_0), ..., \vartheta_k(\mathbf{l}_k))$, where $\vartheta_i$ is the corresponding transform for each $\Theta_i$.

### 3.4.3 Image translation module

The network $\phi : \mathbb{R}^{3+k\times h\times w} \rightarrow \mathbb{R}^{3\times h\times w}$ receives a different frame $\mathbf{f}_{t+k}$ along with the transformed template $\mathbf{T}'$ concatenated along the channel dimension and outputs another image $\tilde{\mathbf{f}}_t$ following an encoder-decoder pathway. Intuitively, this network learns to implicitly disentangle content (*i.e.*, person identity) and pose in an image and transforms the content according to the conditioning pose.

### 3.4.4 Training objectives

Our training objective is split into three parts: A reconstruction loss is associated with the image translation module while a boundary and an anchor loss guide the training of the pose extractor network.

**Reconstruction loss** The main loss component in our method is the reconstruction objective between the original frame $\mathbf{f}_t$ and the output of the image translation network $\phi$. Similar to [Jakab et al., 2020b], we found it to be beneficial to use a perceptual loss with a pretrained VGG network in order to stabilize training:

$$\mathcal{L}_{recon} = ||\psi(\tilde{\mathbf{f}}_t) - \psi(\mathbf{f_t})||_1^1, \tag{3.4}$$

where $\psi(\mathbf{f})$ are feature vectors extracted from a frame by the VGG network and $\tilde{\mathbf{f}}_t$ the output of the image translation network.

**Anchor-point Loss** Each anchor point is being transformed by the corresponding body part transform: $\tilde{\mathbf{a}}_i^j = \Theta_j \mathbf{a}_i^j$. In order to enforce connectivity between body parts, we define a set $A = \{(\tilde{\mathbf{a}}_j^l, \tilde{\mathbf{a}}_k^m) | l \neq m\}$ containing all pairs of transformed anchor points which we require to have the same position. Lastly we compute the mean squared L2 distance of the two points over all tuples:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{M} \sum_{\tilde{\mathbf{a}}_j^l, \tilde{\mathbf{a}}_k^m \in A} ||\tilde{\mathbf{a}}_j^l - \tilde{\mathbf{a}}_k^m||_2^2, \tag{3.5}$$

where $M$ = number of anchor point pairs.

**Boundary loss** Additionally, we found it beneficial for convergence and to prevent the network from outputting transformation parameters leading to a translation of the template shapes out of the image boundaries to enforce the anchor points to be contained within the image:

$$\mathcal{L}_{bx} = \begin{cases} |a_{i,x}^j|, & \text{if } |a_{i,x}^j| > B \\ 0, & \text{otherwise,} \end{cases} \tag{3.6}$$

where $a_{i,x}^j$ is the x-coordinate of the anchor point in pixel space and B corresponds to the size of the (quadratic) image. The loss $\mathcal{L}_{by}$ for the y-component is analogous.

**Combined Loss Formulation** In summary our training objective is expressed as

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{anchor}} + \lambda_2 (\mathcal{L}_{bx} + \mathcal{L}_{by}). \tag{3.7}$$

**Training** Our whole model including all subnetworks is trained in an end-to-end manner. For sampling pairs, we define a course grid and corresponding bounding boxes based on the supplied masks and select frames from the same bounding box to ensure a static background. Note, that the hyper-parameters $\lambda_1, \lambda_2$ are empirically tuned.

## 3.5 Experiments

We evaluate our approach on two different datasets, Human3.6m and an infant dataset from clinical practice. We compare our method to results with various degrees of supervision, as reported in recent literature. Additionally, we performed an ablation study to demonstrate the effectiveness of our proposed anchor and boundary losses.

**Human3.6m** is an industry standard dataset for human pose estimation [Ionescu et al., 2014a] containing 3.6 million images with corresponding 3D and 2D landmarks and captured with different actors in a controlled studio environment with a static background.

**Infant Dataset** A proprietary clinical dataset of videos of 24 moving infants captured in their cod. Details are provided in section 3.3.

**Moving infants in RGB-D (MINI-RGBD)** consists of 12 synthetically rendered movement sequences of infants with different body shapes and backgrounds introduced in [Hesse et al., 2018a]. This public dataset was designed to cover challenging and diverse movements and overall contains 12k video frames together with various label information including landmarks and masks. We introduce this dataset to demonstrate the issue of domain shift when

using these labels as an additional prior when making quantitative comparisons to [Jakab et al., 2020b].

### 3.5.1 Evaluation procedure

**Human 3.6m** In order to make our results comparable to recent literature with regards to the Human3.6m dataset, we adopt the evaluation strategy from [Lorenz et al., 2019b, Zhang et al., 2018b]. The model is trained on subjects 1,5,6,7,8 and tested on subject 9. We restrict the performed actions to {direction, discussion, posing, waiting, greeting, walking} during testing in order to ensure mostly upright poses, resulting in a total of 80k images for testing. 16 landmarks are predicted and compared with the ground truth. More results can be found in the supplemental material. The comparison methods make use of the supplied estimated rough person masks and substract the background. We did not perform this extra preprocessing step.

Many proposed unsupervised or weakly supervised methods require an additional post-processing step and direct supervision via ground truth 2D keypoints. Most use a linear regressor, in order to output human-interpretable landmarks. Our approach returns interpretable landmarks in form of anchorpoints *by design*. These anchorpoints also coincide with the most commonly used landmark definitions, since they are placed around the joints that act as connectors between limbs.

**Clinical and MINI-RGBD infant datasets** As an upper-bound, we fine-tune a fully supervised ResNet based network [Xiao et al., 2018b] pretrained on both ImageNet and adult poses from MPII Human Pose [Andriluka et al., 2014] with roughly 2000 labelled frames from 20 of our 24 subjects from the clinical infant dataset. All self-supervised methods that do not require paired labels during training are trained on all video frames from the 20 subjects of the train split and are evaluated on the remaining four of the test split, resulting in a total of 290k unlabelled training and 471 manually labelled test images. For further comparison, we implemented and trained the method proposed by [Jakab et al., 2020b] on our clinical infant images combined with unpaired poses from the synthetic MINI-RGBD infant dataset [Hesse

et al., 2018a] in order to demonstrate the method's susceptibility to domain shift. 12 landmarks are predicted and compared with the ground truth.

### 3.5.2 Results

**Human 3.6m** For adult pose estimation, we summarise our results in Table 3.2. As expected, none of the self-supervised methods perform as well as the supervised upper bound due to the lack of labels. All prior work makes use of paired or unpaired manual annotations in some capacity. Despite the complete lack of such annotations, our method performs competitively on the same task while only requiring the template. Our model is also able to predict landmarks for difficult poses such as sitting on a chair, as can be seen in Figure 3.4. Errors are largest when self-occlusion occurs. In the most severe cases, when a person is facing away from the camera and both arms and hands are covered by the body, estimating these landmarks becomes extremely difficult without further supervision in the form of manually annotated examples or additional views captured by a second camera.

**Infant pose estimation** The results for infant pose estimation are summarised in Table 3.3 and 3.5. Figure 3.6 displays predictions of our model on infants. Our implemented version of [Jakab et al., 2020b] performs worse despite the access to 11,000 unpaired landmark annotations. We attribute this to domain shift. Since the model relies on adversarial training on these labels, the performance will drop if the latter are not covering a diverse enough range of possible poses. In fact, the authors demonstrate a drop in performance in their own experiments when using labels from a different dataset.

Our model is capable of predicting consistent and interpretable landmarks from images of infants with different body shapes and in different poses. Again, the largest errors are introduced by self-occlusion, especially when arms are positioned in front of the chest and shoulders. This is consistent with our observation that the landmark detection works best on the legs, which are most of the time not positioned above or below other body parts.

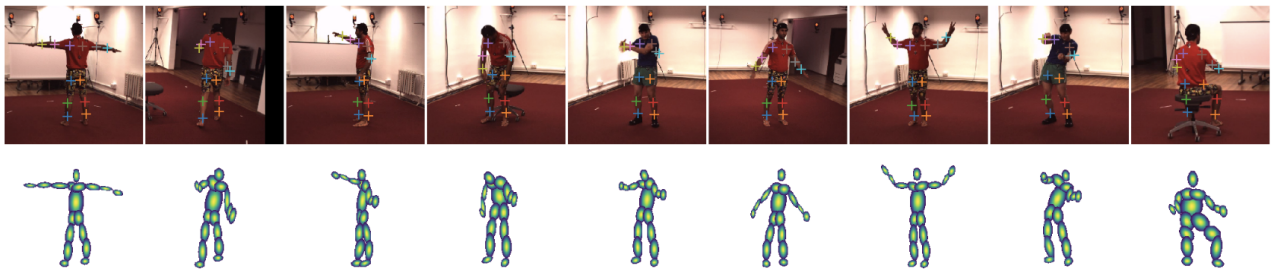


Figure 3.4: Results of our model on random frames from subjects 9 and 11 in Human 3.6m. Top row: input images with detected keypoints. Bottom row: corresponding deformed shape templates

| | BBox | SPP | UL | T |
|---|---|---|---|---|
| Lorenz [Lorenz et al., 2019b] | ✓ | ✓ | ✗ | ✗ |
| Zhang [Zhang et al., 2018b] | ✓ | ✓ | ✗ | ✗ |
| Jakab [Jakab et al., 2020b] | ✓ | ✗ | ✓ | ✗ |
| Ours | ✓ | ✗ | ✗ | ✓ |

Table 3.1: Comparing different sources of labels used for different methods. **BBox**: bounding box centered around the person, **SPP**: supervised post-processing. A linear regressor is applied to map discovered landmarks to human-interpretable locations. **UL**: Use of unpaired manually annotated poses as a prior. **T**: the manual design of a single template including individual body parts and anchorpoints

For a fair comparison of the results, we make this explicit by adding *supervised post-processing* in parenthesis for all results where this step was included. We also compiled Table 3.1 to clearly indicate the different sources of labels that are used by the different methods.

**Ablation study** In order to verify the individual contribution of our loss components, we partially train several models with different loss configurations and present the results in Table 3.4.

**Image translation** Since our decoder network is trained to synthesise images conditioned on the pose input, we are able to perform image editing by combining a frame $\mathbf{f}$ with a different set of poses $p$. The images displayed on the right side of Figure 3.5 illustrate the principle and demonstrate disentanglement between pose and appearance. Moreover, the image translation modules' ability to produce images given a conditioning pose could serve as a tool in the analysis for indicators of impaired neurological development. This, however, remains a topic of future work.

| H36M | all | wait | pose | greet | direct | discuss | walk |
|---|---|---|---|---|---|---|---|
| ***fully supervised upper bound*** | | | | | | | |
| Newell [Newell et al., 2016b] | 2.16 | 1.88 | 1.92 | 2.15 | 1.62 | 1.88 | 2.21 |
| ***self-supervised + supervised post-processing*** | | | | | | | |
| Thewlis [Thewlis et al., 2017b] | 7.51 | 7.54 | 8.56 | 7.26 | 6.47 | 7.93 | 5.40 |
| Zhang [Zhang et al., 2018b] | 4.14 | 5.01 | 4.61 | 4.76 | 4.45 | 4.91 | 4.61 |
| Lorenz [Lorenz et al., 2019b] | 2.79 | – | – | – | – | – | – |
| ***self-supervised (unpaired labels)*** | | | | | | | |
| Jakab [Jakab et al., 2020b] | 2.73 | 2.66 | 2.27 | 2.73 | 2.35 | 2.35 | 4.00 |
| ***self-supervised (template, no labels or post-processing)*** | | | | | | | |
| Ours | 3.31 | 3.51 | 3.28 | 3.50 | 3.03 | 2.97 | 3.55 |

Table 3.2: Comparison with related methods for human landmark detection on Human3.6M; %-pixel distance from ground truth normalized by image size is reported on a per action basis.

| Infants | all | hips | knees | feet | shoulders | hands | params |
|---|---|---|---|---|---|---|---|
| ***fully supervised (fine-tuned) upper bound*** | | | | | | | |
| Xiao [Xiao et al., 2018b] | 1.74 | 2.39 | 1.50 | 1.47 | 1.76 | 1.59 | 34.0 M |
| ***self-supervised (unpaired labels)*** | | | | | | | |
| Jakab [Jakab et al., 2020b] | 8.98 | 6.89 | 8.18 | 13.15 | 5.33 | 11.36 | 8.6 M |
| ***self-supervised (template, no labels)*** | | | | | | | |
| Ours | 4.86 | 3.79 | 4.60 | 5.53 | 3.19 | 7.21 | 7.8 M |

Table 3.3: Comparison with related methods; %-pixel distance from ground truth normalized by image size is reported on a per body-part basis. Our method outperforms prior approaches for this task because it is not influenced by annotation domain shift.

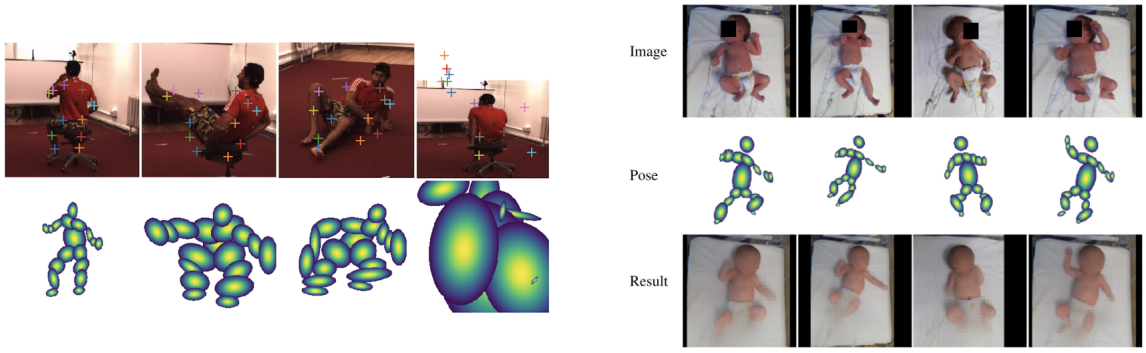



Figure 3.5: Left: Failure cases of model predictions. While sitting is a very difficult pose, facing away from the camera is the most challenging situation resulting in the complete loss of coherent landmarks. Moreover, the model switches left and right when a person is facing away from the camera. Right: Pose conditioning. Top row: Input image; Middle row: New Conditioning Pose; Bottom row: Resulting image. We observe that our image translation submodule is able to disentangle pose and content and produce photo-realistic images with the conditioning poses.

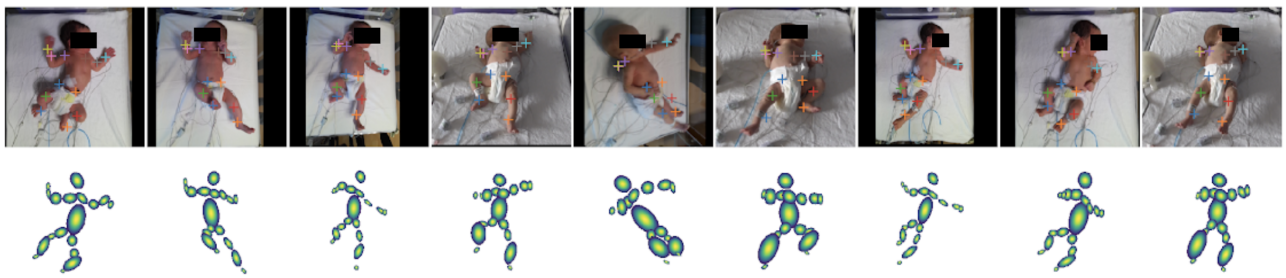


Figure 3.6: Results of our model on random frames from the test subjects of our in-house dataset. Anonymisation was applied to maintain patient privacy. Top row: input images with detected keypoints. Bottom row: corresponding deformed shape templates.

| all | hips | knees | feet | shoulders | hands |
|---|---|---|---|---|---|
| ***anchor and boundary loss (ours)*** | | | | | |
| 6.54 | 3.56 | 5.19 | 8.90 | 4.70 | 7.95 |
| ***with anchor, no boundary loss*** | | | | | |
| 65.99 | 64.65 | 71.70 | 76.33 | 59.00 | 62.59 |
| ***with boundary, no anchor loss*** | | | | | |
| 12.42 | 13.84 | 13.17 | 7.29 | 12.94 | 8.76 |
| ***no anchor, no boundary loss (model diverges)*** | | | | | |
| 446.47 | 433.64 | 385.77 | 662.41 | 502.69 | 276.49 |

Table 3.4: Mean joint distance between ground truth and predicted landmarks in % of image size. All models were trained for 10 epochs.
Without our anchor or boundary loss, the model diverges and all predicted landmarks lie outside of image boundaries (values above 100).

## 3.6 Downstream application study

To study the downstream effects of 2D pose prediction accuracy in a realistic setting we integrate our method into a setup as it is currently used for clinical evaluation of neonatal movement quality. Physiological movement quality assessment requires 3D coordinates, thus a fully supervised 3D lifting network is applied after 2D pose estimation. The lifting model is adapted from the work of Martinez et al. [Martinez et al., 2017b]. This model was first pre-trained on the MPI_INF_3DHP adult dataset [Mehta et al., 2017] before fine-tuning on the infant MINI-RGBD dataset [Hesse et al., 2018a]. Figure 3.7 compares the the qualitative differences when results from previous 2D pose estimation methods [Jakab et al., 2020b] are used or ours. Quantitatively this can be captured by evaluating the joint position error in mm. These results are

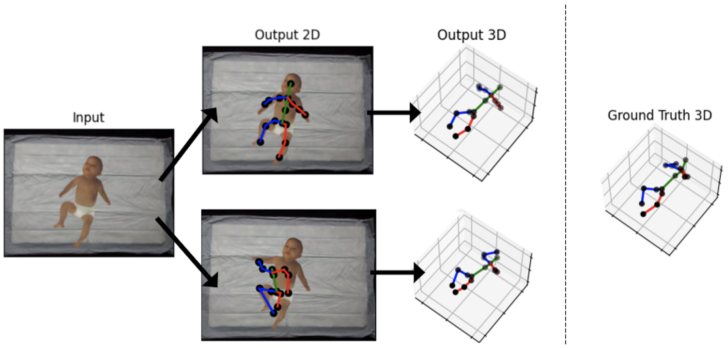



Figure 3.7: Visual results from the lifting network on a synthetic infant image. Top: ours, bottom: [Jakab et al., 2020b]. The accuracy of the predicted 2D keypoints has a noticeable impact on the final 3D predictions.

| Synthetic Infants | all | hips | knees | feet | shoulders | hands |
|---|---|---|---|---|---|---|
| ***self-supervised (unpaired labels)*** | | | | | | |
| Jakab [Jakab et al., 2020b] | 59.7 | 10.1 | 58.4 | 73.3 | 58.0 | 101.5 |
| ***self-supervised (template, no labels)*** | | | | | | |
| Ours | 44.7 | 8.0 | 57.5 | 80.9 | 35.6 | 78.7 |

Table 3.5: 3D lifting results. The error is reported as the distance between predicted and ground truth 3D landmarks in mm.

summarised in Table 3.5. Our method performs better compared to [Jakab et al., 2020b] in all landmarks except feet. Since the 3D-lifting can not correct for inaccuracies in the 2D landmark estimation, the larger error as shown in table 3.3.

## 3.7 Summary

We presented a novel approach to unsupervised human pose estimation by transforming a part-based shape template. Given a video of a person moving in front of a static background, we are able to predict human interpretable keypoints without requiring any paired or unpaired labels. We have exhibited our method's performance on two different datasets and achieved similar or better results in unsupervised human pose estimation while only requiring a simple canonical 2D pose template instead of numerous manual labels. Moreover, our method is able to adapt to humans or infants of different shapes and sizes and alternative keypoints can be defined easily by choosing different locations on each body part. We also demonstrate an increase

in performance when using our method's extracted poses in downstream tasks like 3D pose extrapolation, which is of significance for both industrial and medical applications. Although some manual work is required to define the template, this could be easily realised in a simple GUI, where the user would be able to draw body parts and define anchor- as well as keypoints within minutes. Moreover, our method can generate realistic images conditioned on pose once trained, opening up the possibility for other applications including pose transfer from a domain where labels are readily available, to another where labels are limited.

**Limitations** Our method is limited by occlusions and partial views of the body and is currently not able to distinguish back-facing from front-facing poses. If not accounted for, this can result in large errors due to the switching of joints with their left/right counterpart. The problem only occurs for adult pose estimation, as infants are lying on their backs during a recording and are physically not able to turn around. We believe that the inclusion of a directional vector could alleviate the observed limitation.
The lifting network on the downstream application study maps 2D coordinates to 3D and is fully supervised with synthetic data. Naturally, the question arises whether our method could be extended to predict 3D pose directly in an unsupervised fashion and is answered in the next chapter.

## 3.8 Appendix

The following paragraphs give more detailed information about network architectures and training hyperparameters, as well as additional qualitative results.

### 3.8.1 Network architectures

In this section, we present the details regarding network architecture. Each block represents a layer. Abbreviations are as following: k=kernel size, s=stride, d = dilation, ch= number of channels, n=number of (fully connected) neurons.

| | |
|---|---|
| Conv (k=3, s=1, d=1, ch=32) | 256x256 |
| Conv (k=3, s=1, d=4, ch=32) | |
| Conv (k=3, s=2, d=1, ch=64) | 128x128 |
| Conv (k=3, s=1, d=8, ch=64) | |
| Conv (k=3, s=2, d=1, ch=128) | 64x64 |
| Conv (k=3, s=1, d=1, ch=128) | |
| Conv (k=3, s=2, d=1, ch=256) | 32x32 |
| Conv (k=3, s=1, d=1, ch=256) | |
| Conv (k=3, s=2, d=1, ch=32) | 16x16 |
| Conv (k=3, s=1, d=1, ch=256) | |
| Conv (k=3, s=2, d=1, ch=256) | 8x8 |
| Conv (k=3, s=1, d=1, ch=256) | |
| Conv (k=3, s=2, d=1, ch=256) | 4x4 |
| Conv (k=3, s=1, d=1, ch=256) | |
| Linear (n=512) | 6k |
| Linear (n=6k) | |

Figure 3.8: Architecture of **pose extractor** $\varphi$. All layers are followed by batch normalisation and LeakyReLU except for the last.

| | |
|---|---|
| Conv (k=3, s=1, d=1, ch=32) | 256x256 |
| Conv (k=3, s=1, d=4, ch=32) | |
| Conv (k=3, s=2, d=1, ch=64) | 128x128 |
| Conv (k=3, s=1, d=8, ch=64) | |
| Conv (k=3, s=2, d=1, ch=128) | 64x64 |
| Conv (k=3, s=1, d=1, ch=128) | |
| Conv (k=3, s=2, d=1, ch=256) | 32x32 |
| Conv (k=3, s=1, d=1, ch=256) | |
| Bilinear Upsampling | 64x64 |
| Conv (k=3, s=1, d=1, ch=128 | |
| Bilinear Upsampling | 128x128 |
| Conv (k=3, s=1, d=1, ch=32 | |
| Bilinear Upsampling | 256x256 |
| Conv (k=3, s=1, d=1, ch=32 | |
| Conv (k=3, s=1, d=1, ch=3 | |

Figure 3.9: Architecture of **image decoder** $\phi$. All layers are followed by batch normalisation and LeakyReLU except for the last.

### 3.8.2 Additional qualitative results

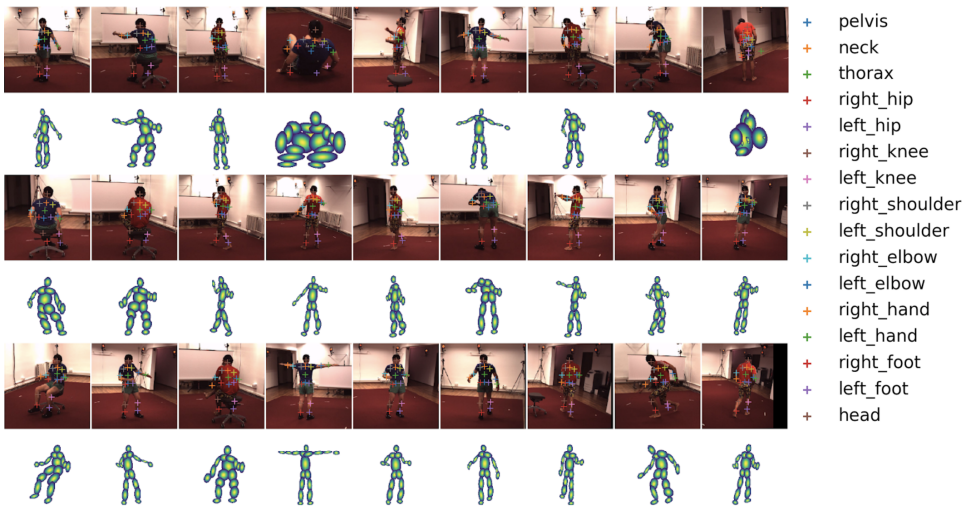



Figure 3.10: Results for Human 3.6m. Markers denote detected landmarks.




Figure 3.11: Results for infants. Markers denote detected landmarks.

## 3.9 Training and evaluation details

### 3.9.1 Human3.6m

The model is trained with the Adam optimizer [Kingma and Ba, 2014], a learning rate of 1e-4 and a batch size of 48. For evaluation, we choose 16 keypoints corresponding to following indices: 0, 13, 12, 1, 6, 2, 7, 25, 17, 26, 18, 27, 19, 3, 8, 14. While most of these keypoints conincide with our anchor points, we place two additional points on the thorax and the head in our default template.

### 3.9.2 Infants

For the infants, both models are trained the Adam optimizer [Kingma and Ba, 2014], a learning rate of 1e-4 and a batch size of 48. In order to train [Jakab et al., 2020b] with labels from the synthetic infants dataset, we project 3d coordinates into the image plane via a perspective camera transformation. Since the camera position in our clinical infant dataset is varying in terms of tilting angles, we augment the data via random rotations of the camera before projection in an effort to mimic these conditions.

# Chapter 4

# 3D-Pose Estimation via Neural Rendering

## 4.1 Introduction

In the previous chapter we presented a method for self-supervised 2D-pose estimation via differentiable shape templates. It is a natural question to ask whether the approach can be extended to 3D as well. Estimating 3D-pose from a single camera, however, is a lot more challenging, even more so if we restrict ourselves to the self-supervised setting. As outlined in section 2.6.2, in recent years combining fully-differentiable rendering pipelines with the representational power of deep neural networks has become a very popular and active field of research in computer vision for novel-view synthesis, digital avatars and many other applications. Combining computer-graphics rendering pipelines with neural networks in various settings such as controlling the elements of a 3D scene or the 3D transformation of existing objects has been coined *neural rendering*. This chapter is based on [Schmidtke et al., 2022] and is an extension of the approach introduced in the previous chapter from 2D to 3D pose estimation. Differentiable rendering techniques in combination with a 3D shape template will be utilised to train the system end-to-end with self-supervision.

## 4.2 Contextual overview

Inferring 3D properties of our world from 2D images is an intriguing open problem in computer vision, even more so when no direct supervision is provided in the form of labels. Although this problem is inherently ill-posed, humans are able to derive accurate depth estimates, even when their vision is impaired, from motion cues and semantic prior knowledge about the perceived world around them. This is especially true for human pose estimation. As demonstrated in the previous chapter, self-supervised learning can be an effective technique to utilise large amounts of unlabelled video and image sources and provide flexibility in special domains such as medical imaging where data is scarce. In this chapter we therefore focus on self-supervised 3D pose estimation from monocular video, a key element of a wide range of applications including motion capture, visual surveillance or gait analysis.

Methodologically, we borrow the self-supervision framework from the previous chapter, where we model pose as a factor of variation throughout different frames of a video of a single human and a static background. More formally, self-supervision is provided by formulating a conditional image reconstruction task: given a pose input different from the current image, what would that image look like if we condition it on the given pose? This time we choose to represent pose as a 3D template consisting of connected parts which we transform and project to two-dimensional image space, thereby inferring 3D pose from monocular images without explicit supervision.

More specifically, our method builds upon the recent emergence and success of combining deep neural networks with an explicit 3D to 2D image formation process through fully differentiable rendering pipelines. This inverse-graphics approach follows the analysis by synthesis principle of generative models in a broader context: We hope to extract information about the 3D properties of objects in our world by trying to recreate their perceived appearance on 2D images. Popular rendering techniques rely on different representations including meshes and polygons, point clouds or implicit surfaces. In our work we make use of volume rendering with a simple occupancy function or density combined with a texture field that assign an occupancy between $[0, 1]$ and RGB colour value $c \in \mathbb{R}^3$ for every point defined on a regular 3D grid.

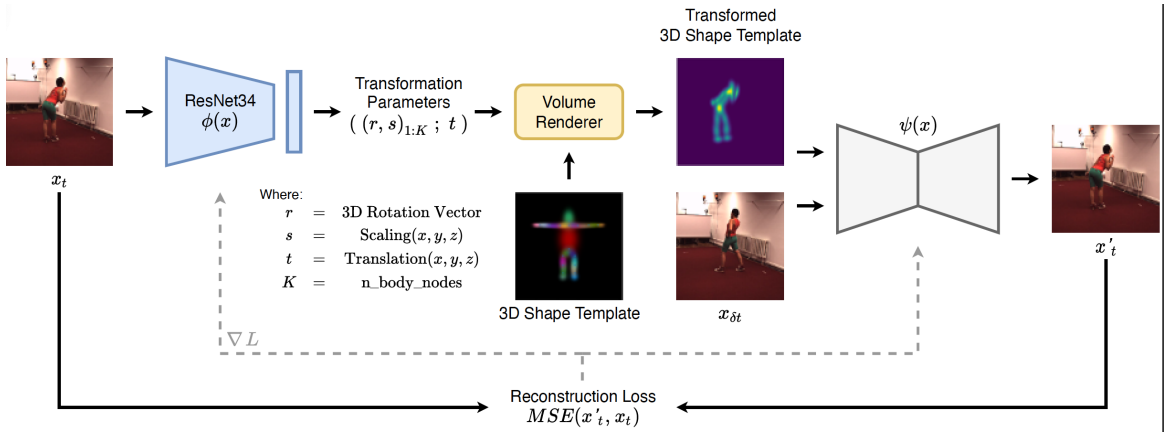



Figure 4.1: Overview of our method. Analogous to Figure 3.3 in the previous chapter, except our pose template is made up of 3D voxels and projected onto 2D via volume-rendering. The whole approach is trained end-to-end, meaning gradients are flowing backwards through the rendering step.

## 4.3 Method

Our approach relies on self-supervision through image reconstruction conditioned on a transformed and rendered shape template. The images are sampled from a video containing a single person moving in front of a static background. More formally, the goal is to reconstruct a number of frames $(\mathbf{x}_{t_1}, \mathbf{x}_{t_n})$ from random time points $t_1, ..., t_n$ in a video with access to *one* frame $\mathbf{x}_{t_k}$, again sampled randomly, and rendered images of transformed templates $\mathbf{T_1}, ..., \mathbf{T_n}$.

Our method can be viewed in two distinct steps; regression of template transformation parameters, and image reconstruction, where both steps are parameterized using deep convolutional neural networks. An encoder network $\phi$ regresses rotation, translation and scale parameters from frame $\mathbf{x}_t$ in order to transform each skeletal node of a 3D shape template, $\mathbf{T}$. The generator network $\psi$ takes as input; (a) a frame $\mathbf{x}_{\delta t}$, from a different time instance of the same sequence where the same person assumes a different pose, and (b) a rendered image of the transformed 3D template while being tasked to reconstruct frame $\mathbf{x}_t$.

The encoder $\phi$ consists of a convolutional neural network for feature extraction followed by a number of linear layers and a reshape operation. The generative network $\psi$ resembles a typical convolutional encoder-decoder structure utilised for image translation, where feature maps are subsequently downsampled via strided convolutions and the number of features increases.

For the decoder we utilise bilinear upsampling and spatially adaptive instance normalisation (SPADE) (see 2.1.4) to facilitate semantic inpainting of the rendered template image.

### 4.3.1 Template and volume rendering

#### Shape Template

The shape template, $\mathbf{T}$, consists of $K$ Gaussian ellipsoids that are arranged in the shape of a human. Each skeletal node, denoted as $\mathbf{T}_k$, is defined on a regular volumetric grid and represents a single body part. All ellipsoids are parameterised by their mean $\mu_k$ and co-variance $\mathbf{\Sigma}_k$. On the volumetric grid we define two functions: a scalar field $f : \mathbb{R}^3 \rightarrow [0, 1]$ that assigns a value to each point $(x, y, z)$ on the grid — in the volume rendering literature it is commonly referred to as the occupancy function, and a vector field $c : \mathbb{R}^3 \rightarrow \mathcal{C} \subset \mathbb{R}^3$ specifying the RGB-colour for each point, commonly referred to as the colouring function. We use uniform colours for each part and define the opacity via the 3D Gaussian ellipsoid, *i.e* the density decreases exponentially moving outward from the centre of the blob.

#### Raycasting and emission absorption function

We make use of an existing implementation of the raycasting algorithm shipped with the PyTorch3D package [Ravi et al., 2020] to render the template image. Given a camera location $\mathbf{r}_0 \in \mathbb{R}^3$, rays are "emitted" from $\mathbf{r}_0$ that pass through each pixel $\mathbf{u}_i \in \mathbb{R}^3$ lying on a 2D view plane $\mathcal{S}$ by sampling uniformly spaced points along each ray starting from the intersecting pixel:

$$\mathbf{p}_j = \mathbf{u}_i + j\delta s, \tag{4.1}$$

where $j$ is the step and $\delta s$ the step size that depends on the maximum depth and number of points along each ray.

The colour value at each pixel location $\mathbf{u}_i$ is then determined by a weighted sum of all colour

values of the points sampled along the ray:

$$\mathbf{c}_i = \sum_{j=0}^{J} w_j \mathbf{c}_j \tag{4.2}$$

The weights $w_j$ are computed by multiplying the occupancy function $f(x)$ with the transmission function T(x) evaluated at each point $\mathbf{p}$ along the ray:

$$w_j = f(\mathbf{p}_j) \cdot T(\mathbf{p}_j), \tag{4.3}$$

where $T(\mathbf{x})$ can be interpreted as the probability that a given ray is not terminated, *i.e.* fully absorbed, at a given point $\mathbf{x}$ and is computed as the cumulative product of the complement of the occupancy function of all $k$ points up until $\mathbf{p}_j$:

$$T(\mathbf{x}) = \prod_{k} (1 - f(\mathbf{x}_k)) \tag{4.4}$$

Repeating this for all pixels in the view plane results in a 2D projection of our 3D object representing the rendered image $\mathbf{f}_r \in \mathbb{R}^{3\times h\times w}$.

### 4.3.2 Pose regression and shape transformation

In order to estimate the skeletal pose of a given frame, we use the encoder network, $\phi : \mathbb{R}^{3\times h\times w} \rightarrow \mathbb{R}^{3K+3}$ based on the ResNet-34 architecture [He et al., 2016a].

The encoder maps a colour input image of size $h \times w$ to $K$ rotation and scale vectors, $(\mathbf{r}, \mathbf{s})_{1:K} \in \mathbb{R}^3$, and a single global translation vector $\mathbf{t} \in \mathbb{R}^3$ for the camera. $K$ denotes the number of transformable parts in the template. Here, rotation is parameterised via axis-angle representation, and is subsequently converted to 3D transformation matrices using the Rodrigues' rotation formula. Combined with the scaling parameter for each axis, the resulting matrix defines the affine mapping, excluding the sheer component, for spatial transformation of each skeletal node.

After construction of the 3D transformation matrix, each Gaussian ellipsoid of the template

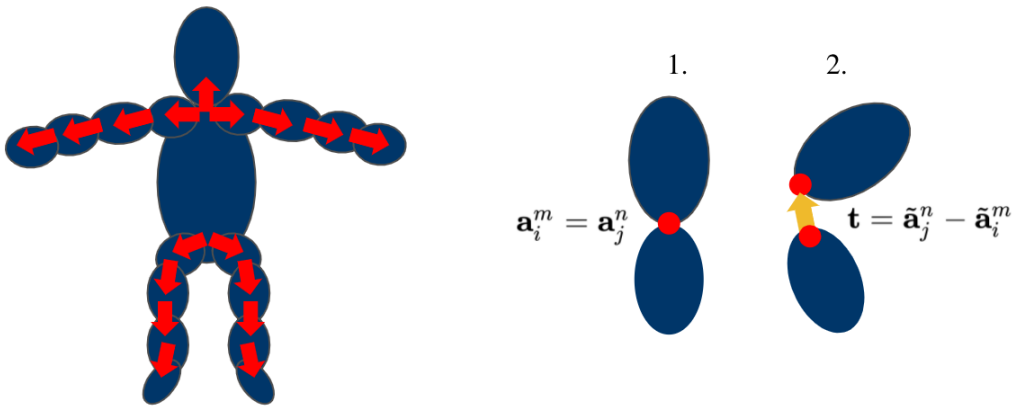



Figure 4.2: Illustration of the kinematic chain. Red circles denote anchor-points. Following a transformation of the upper part, the translation $\mathbf{t} = \tilde{\mathbf{a}}_j^n - \tilde{\mathbf{a}}_i^m$ is applied to enforce continuity

$\mathbf{T}$, with occupancy $f_k(\mathbf{x})$ and colour field $c_k(\mathbf{x})$, gets transformed according to the regressed parameters. Finally, utilising the aforementioned ray-tracing method we render an image based upon our transformed template by summing together all transformed occupancy and colour fields and clipping to a maximum value of 1:

$$\tilde{\mathbf{T}}_k = \Omega_k(\mathbf{T}_k) \tag{4.5}$$

$$\mathbf{f}_r = \mathcal{R}\big(\sum_k f_k(\mathbf{x}), \sum c_k(\mathbf{x})\big), \tag{4.6}$$

,where $\mathcal{R}$ denotes the rendering operation.

### 4.3.3 Kinematic chain

Instead of relying on an additional loss to enforce connectivity between body parts as in the previous chapter in the form of the anchor-point loss, we define a kinematic chain along which each body part is reconnected to its parent via a translation after rotation and scale have been applied. This removes a redundant optimisation objective and increases robustness during training.

Given a parent and child body part with indices $n$ and $m$ respectively, we define anchor-points $\mathbf{a}_i^n$ and $\mathbf{a}_j^m \in \mathbb{R}^3$ on each part representing the area of overlap in the non-transformed template (see Figure 4.2 right). If body part $m$ is being transformed, the position of the anchor point

changes: $\tilde{\mathbf{a}}_j^m = H\mathbf{a}_j^m$, where $H \in \mathbb{R}^{3\times3}$ specifies a transformation matrix. To ensure continuous connectivity, we apply the transformation for the child body part in an analogous way, it is reconnected with the parent node by applying translation $\mathbf{t} = \tilde{\mathbf{a}}_j^n - \tilde{\mathbf{a}}_i^m$. We first transform the core, and then proceed with all other parts in an iterative fashion along the kinematic chain as depicted in Figure 4.2 left.

### 4.3.4 Loss function

The loss function is a sum of individual components: the reconstruction loss as the pixel-wise $l^2$-norm between the decoder output and the original image and a boundary loss of the form

$$\mathcal{L}_{bx}^i = \begin{cases} |\hat{a}_{x,i}|, & \text{if } |\hat{a}_{x,i}| > 1 \\ 0, & \text{otherwise} \end{cases} \qquad \mathcal{L}_{bx} = \sum_i \mathcal{L}_{bx}^i \qquad \mathcal{L}_b = \mathcal{L}_{bx} + \mathcal{L}_{by} \tag{4.7}$$

where $\hat{a}_{x,i}$ is the x-component of a projected and transformed anchor point. Note that we normalise image coordinates to $(-1, 1)$.

We also regularise the pose regression via the $l^2$-norm of the rotation vector $\mathbf{r}$ and decay this term linearly to 0 after 500 iterations. Overall, our objective function is:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \mathcal{L}_b + \alpha * \sum_K ||\mathbf{r}_k||_2, \quad \alpha = \min(1 - 0.02 * \text{iter}) \tag{4.8}$$

### 4.3.5 Infant-adjusted SMPL model

We also experimented with an alternative representation provided by a SMPL [Loper et al., 2015c] body model adjusted to infants from [Hesse et al., 2018b]. In SMPL, a mesh template resembling a generic adult human is transformed via identity-based shape parameters $\boldsymbol{\beta}$ and skeleton-pose depended parameters $\boldsymbol{\theta}$. [Hesse et al., 2018b] create a new template based on infant body-proportions and learn a new model, termed Skinned Multi-Infant Linear model (SMIL), by registering the template with a scaled down version of the original pose blend shapes of SMPL with 3D-point clouds obtained from filming a total of 37 moving infants with

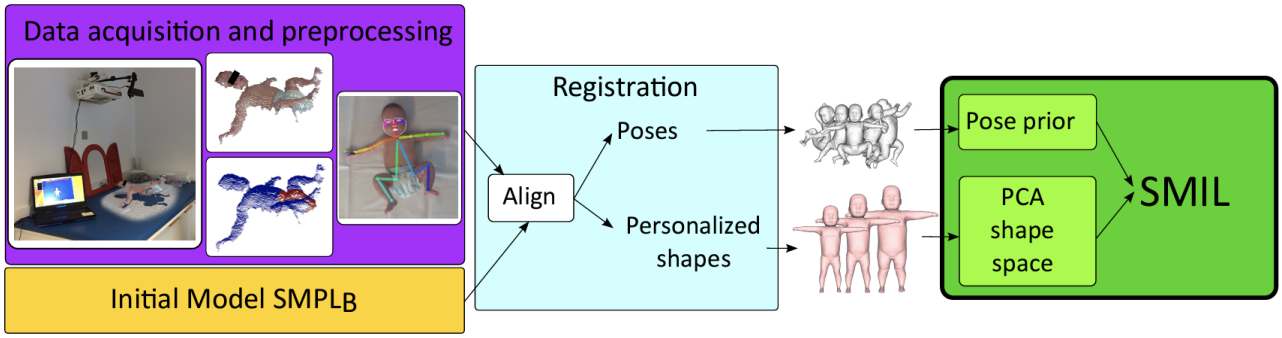



Figure 4.3: Excerpt from [Hesse et al., 2018b]. An infant-adjusted SMPL body model [Loper et al., 2015b] is registered to 3D point clouds of moving infants acquired via a depth sensor. A new shape and pose space is learnt via PCA and a prior in the form manual 2D-pose labels.

an RGB-D camera. Subsequently, the new model is learned by performing PCA on the unposed shapes for each infant and incorporating the prior from roughly 37,000 manual 2D-pose labels. Figure 4.3 illustrates the process.

### Method

**SMPL parameters** Instead of regressing parameters for a 3D rotation and scaling as described in 4.3.2, we train a CNN $\phi$ based on a ResNet architecture [He et al., 2015] to regress SMPL pose and shape parameters $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and $\boldsymbol{\beta} \in \mathbb{R}^{20}$ from an input image $\mathbf{x}$:

$$\boldsymbol{\theta}, \boldsymbol{\beta} = \phi(\mathbf{x}) \tag{4.9}$$

**Differentiable rendering of meshes** We utilise PyTorch3D [Ravi et al., 2020] which makes rasterisation of meshes differentiable via soft assignments inspired by [Liu et al., 2019]. The rendering pipeline is displayed in Figure 4.4.

## 4.4 Experiments

Analogous to the previous chapter, we train and qualitatively evaluate our model on Human 3.6M [Ionescu et al., 2014b], a motion-capture dataset including 11 actors performing various

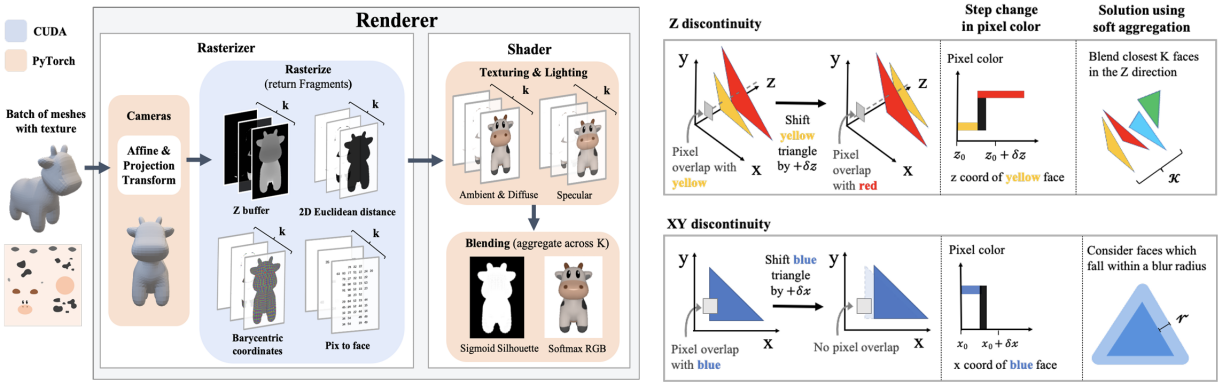



Figure 4.4: Excerpt from [Ravi et al., 2020] displaying the graphics-pipeline. Soft rasterization is introduced to tackle the non-differentiability of rasterisation and z-buffering due to discrete sampling steps.

activities while being filmed by four different cameras in a studio setting. Following [Jakab et al., 2020a] and [Luca Schmidtke and Kainz, 2021], we train on subjects 1, 5, 6, 7, test on 9 and 11 and restrict activities to mostly upright poses, resulting in roughly 700,000 images for training. We sample video frames in pairs containing the same person in different poses, but with the same background utilising bounding boxes derived from the masks and utilise the Adam optimiser [Kingma and Ba, 2015].

As an additional experiment, we also train the volume rendering as well as mesh-based SMPL-model on the infant dataset described in section 3.3, which includes videos of 24 newborn infants. We train on 20 and test on 4 subjects.

## 4.5 Results

We restrict our evaluation to qualitative results for Human3.6m and infants, which are displayed in Figures 4.5 and 4.7, respectively. This is motivated by the fact that a quantitative evaluation would still show large errors due to depth ambiguity and switching of left and right landmark counterparts. We also do not compare to previous work since we are the first to achieve this without any supervision through manual pose landmarks either in 2D or 3D. The purpose if this work is to demonstrate that the concept of self-supervision through conditional image

translation can be extended to 3D pose estimation. A simple image reconstruction loss serves as the sole basis for a learning signal, and is properly propagated back through the rendering pipeline to the weights of the neural network that regresses the transformation parameters of our 3D human model. However, there are several issues that still need to be solved: The model is currently not able to distinguish left and right, as can be observed in Figure 4.5 (fourth image from the right), where the subject is facing away from the camera, but the template remains in the front-facing configuration. The model also mostly generates limbs facing away from the camera (third image from the right). We hypothesise that due to depth ambiguity in 2D and limitations in pose variety due to the restricted sampling the decoder can perfectly reconstruct the image despite the wrong orientation of the limb. For the infants in Figure 4.7, we observe that the arms, additionally to facing away from the camera, do not align with the real images even in the 2d-projection. Interestingly, the legs seem to be better aligned. Our hypothesis is that infants display a lot more movement, and therefore variety, with their legs rather than their arms. Due to our training regime, the decoder network might not require spatially aligned arms in order to reconstruct the image faithfully.

Figure 4.6 displays qualitative results for the infant-adjusted SMPL model. We were not able to regress meaningful shape and pose parameters within our self-supervised training regime. These findings are in line with other research works such as [Wang et al., 2022] noting the reliance of explicit, surface-based body-models on more than pure image supervision.

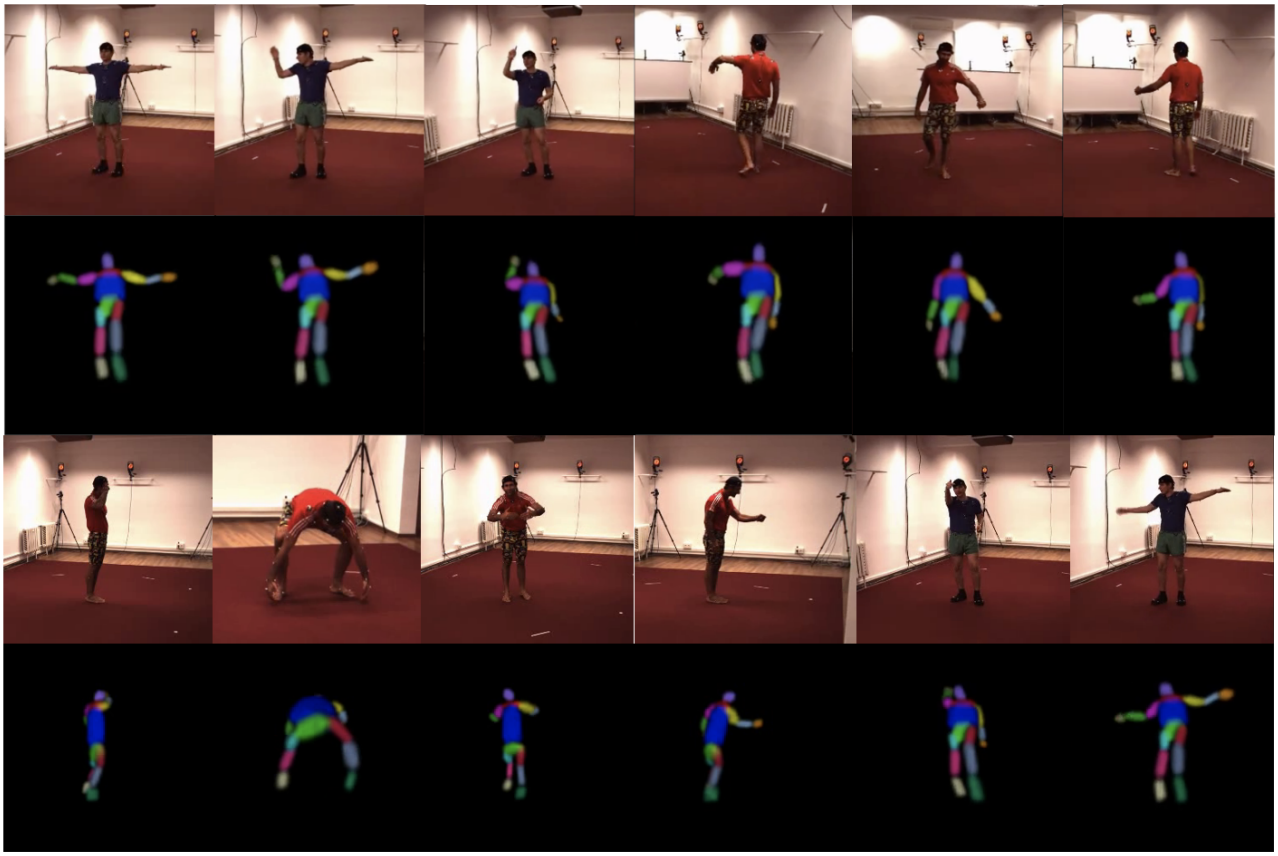


Figure 4.5: Results on the two evaluation subjects. Top row: input image. Bottom row: predicted pose in the form of a transformed and rendered 3d shape template. Different colours denote different body-parts.



Figure 4.6: Left: rendered image from the baby-model in the standard t-pose. Middle: input image. Right: rendered image from regressed pose and shape SMPL parameters.

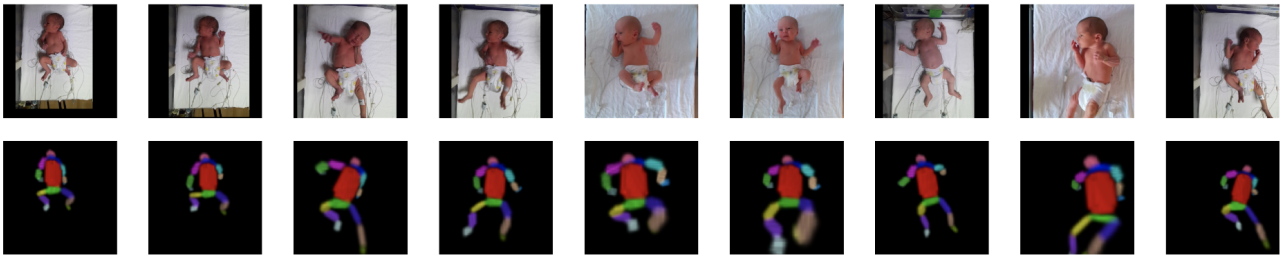


Figure 4.7: Results on the infant dataset described in 3.3. Top row: input images. Bottom row: predicted pose in the form of a transformed and rendered 3d shape template. Different colours denote different body-parts.

## 4.6 Summary

In this chapter we introduced a volume-rendering-based approach that is fully-differentiable by default and found that we can find meaningful pose representations in 3D from monocular images without relying on any landmark labels during training and inference.

In a preliminary experiment, we tested a mesh-based SMPL model in combination with a differentiable soft-rasterisation rendering technique and found that, despite optimising for hyperparameters, the model does not converge to meaningful representations.

In contrast, we have found a voxel-based representation combined with volume-rendering approach to be more suitable for end-to-end training, we hypothesise that this is due to more accurate gradients being propagated to our network weights through this rendering pipeline. Despite issues with depth ambiguity, the qualitative results are encouraging and demonstrate the feasibility of combining differentiable rendering techniques and self-supervision purely based on image reconstruction losses from a single view.

As we shall see in the following chapter, replacing the image translation task with a purely generative approach with separate fore- and background similarly to [Yang et al., 2022] might prove to be very successful in extending the approach to non-static backgrounds as well.

# Chapter 5

# Pose Estimation Via Generative Modelling

## 5.1 Introduction

The methods presented in the previous chapters demonstrate image translation conditioned on shape templates provides a strong self-supervision signal for finding pose landmarks. However, they remain restricted to a studio-type setting because the background needs to remain constant between the two images of the same person in different poses. Arguably, the vast majority of videos available on the internet *do not* fall in this category, as there are either moving objects in the background or the background shifts due to movement of the camera itself. This is especially true for videos recorded through hand-held devices such as smartphones. One viable option would be to first run an off-the-shelf segmentation model to separate fore- from background, but that would defeat the purpose of striving for self-supervision, since these methods have been trained in a fully-supervised way. Moreover, this could lead to a repeated need for dataset-curation and re-training when large domain-shifts lead to worse performance of these models. Generative modelling, more specifically image generation, has made tremendous gains over the past 5 years and provides an intriguing self-supervision signal for a variety of different tasks. The following chapter introduces a new concept for estimating body landmarks: *analysis by*

*synthesis*, or learning through generating, which is argued to be a central element in human recognition [Stevens, 1960, Parr and Friston, 2018]. Intuitively, the brain must have some internal model of the world in order to properly reason about the most likely causes of a certain sensory input. Put into the context of pose estimation, learning a representation of human anatomy as a factor in generating image data of humans might prove helpful in estimating it from existing images. Viewed through this lens, the previous chapters introduced methods that learn by generating image data conditioned on a body model as well, albeit only parts of it. The method in this chapter can be seen as a natural extension in that it aims to include fore- as well as background into the generative process for increased performance and generalisability. We achieve this by designing and training a disentangled generative model that synthesises fore- and background separately. The foreground consists of a human and is itself disentangled into pose conditioned on our already familiar pose template introduced in the earlier chapters and appearance. Crucially, this model can synthesize realistic images accompanied by a body-part segmentation as well as pose landmarks that can be utilised as synthetic training data for a fully-supervised model. We also explore an alternative where pose is learnt by reconstructing images through the lens of a pretrained, disentangled and pose-conditioned generator network.

## 5.2 Contextual overview

We revisit the problem of self-supervised human pose estimation from monocular images, but through the lens of generative models.

Generative models, specifically Generative Adversarial Networks [Goodfellow et al., 2014b, Karras et al., 2019a], and diffusion models [Dombrowski et al., 2022, Rombach et al., 2021] play a prominent role in self-supervised learning due to their ability to approximate the supposed high-dimensional manifold on which the distribution of natural images are believed to lie. These models can infer knowledge about the inherent structure of their input by learning how to generate it, especially if the model is capable of doing so via disentangled and semantically meaningful generative factors. Learning inherent structure in data from generating it forms the

core principle around our proposed method which can be subdivided into two stages: In the first stage we make use of generative adversarial networks that, with appropriate modifications, are capable of composing realistic scenes from several independent factors. In the context of pose-estimation and human parsing, these factors can be defined as the background, the human body in a certain pose configuration, and the general appearance, *i.e.* clothes and texture of the person in the scene. We introduce a fully differentiable and transformable spatial shape template, which resembles connected human body parts. This template can be used for inpainting the foreground image to explicitly drive the model towards treating pose as an independent generative factor. In the second stage, once our initial disentangled generative model has been trained to produce realistic images, we can generate an arbitrarily large amount of these images to train a fully-supervised model on pose-estimation and body-part segmentation or try and reconstruct input images through the lens of the disentangled generative model. In summary, we make the following contributions:

- We learn to estimate 2D pose landmarks as well as part-segmentations from monocolular images by *learning how to generate them.* We achieve this by learning a disentangled generative model that composes images from a background and foreground image conditioned on a transformable shape template.

- This allows us to do unsupervised pose-estimation independent from contextual and background information, thus expediting new avenues for model generalisation and domain adaption.

- We provide a solution to the self-occlusion problem frequently described in weakly- and self-supervised pose estimation and human parsing literature [He et al., 2022, 2021].

- We demonstrate the feasibility of our approach qualitatively and quantitatively on two different datasets including diverse scenarios.

## 5.3 Related work

**Generative models** The quality of generated images has massively improved since the introduction of generative adversarial networks [Goodfellow et al., 2014b, Karras et al., 2019a] and diffusion models [Dhariwal and Nichol, 2021, Ho et al., 2020]. Related to our approach, several works propose to learn disentangled models that are capable of generating one or more foreground objects independent from the background. [Kwak and Zhang, 2016] introduce several generator networks trained adversarially and connected via a recurrent neural network outputting objects that are composited via alpha blending in the final image. [Yang et al., 2017] expand on this by further disentangling foreground generators into *shape* and *appearance* as well as a more elaborate conditioning between objects. [Bielski and Favaro, 2019] make use of small random spatial pertubations of the foreground in the form of translation in order to avoid trivial solutions where either the foreground or background generator synthesises the entire image. [Yang et al., 2022] combine this perturbation with maximising mutual information between foreground and latent code similar to [Singh et al., 2019]. However, none of the above mentioned works provide part based segmentation or even keypoint information. [He et al., 2021] utilise spatially adaptive normalisation (SPADE) (see section 2.1.4 in the background chapter) together with low-frequency positional encoding and a combination of geometric constraints in the objective function. They achieve segmentation into different parts in several object categories, but the partition remains fully flexible and does not necessarily correspond to predefined anatomical locations of the human body.

**Pose and keypoints from shape priors** As outlined in section2.6.1 in the background chapter, the human body can be represented by individual connected parts that can move independently within a certain range. This can be exploited as an inductive bias for pose estimation algorithms. Earlier work utilises this in the form of pictorial structures [Felzenszwalb and Huttenlocher, 2005, Andriluka et al., 2009b, Ouyang et al., 2014, Pishchulin et al., 2013, Ramakrishna et al., 2014, Sapp et al., 2010, Yang and Ramanan, 2011] while more recent works make use of differentiable shape templates [Kundu et al., 2020b, Schmidtke et al., 2021] or other

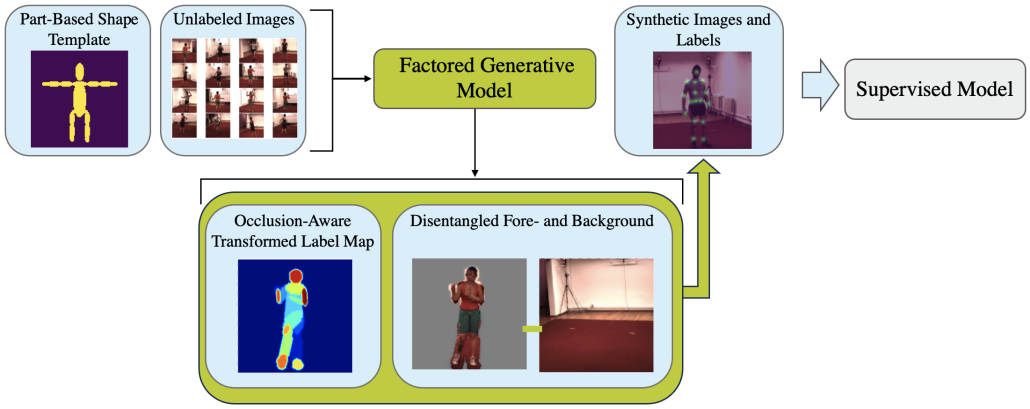



Figure 5.1: We manage to factorise a generative model into pose, person appearance and background via a simple shape template. Large amounts of synthetic data can be created to train a supervised model for pose estimation and human body parsing.

forms of pictorial renditions [Jakab et al., 2019, He et al., 2022] combined with self-supervision objectives, most notably image reconstruction conditioned on a corrupted or altered image together with some form of pose representation. While the former requires either constant background or some form of foreground-background separation in videos, the latter requires linear regression with ground truth keypoints.

## 5.4 Overview

Conceptually our approach can be divided into two distinct stages as illustrated in Fig. 5.1. In the first stage we train a generative model $G$ parameterised by neural networks $\psi_{\mathrm{b}}, \psi_{\mathrm{f}}, \psi_{\mathrm{m}}$ for generating and compositing background, foreground and mask, respectively. The foreground and mask represent humans in our use-case and are conditioned on a spatially transformed shape template consisting of $K$ part maps $\mathbf{T} = \{\mathbf{T}_1, ..., \mathbf{T}_K\}, \mathbf{T_i} \in \mathbb{R}^{h \times w}$. More specifically, $G$ maps a low-dimensional vector $\mathbf{z}$ sampled from a standard normal distribution to a *background* image $\mathbf{b}$, a *foreground* image $\mathbf{f}$ and a *mask* image $\mathbf{m}$ through separate networks. The mask generator maps $\mathbf{z}$ to a pose representation $\mathbf{p}$ represented by a transformed shape template and a resulting mask $\mathbf{m}$. Similarly to the methods introduced in the earlier chapters, the shape template serves as a representational bottleneck for the generator in order to effectively learn pose disentangled from other image content via a pure image generation objective. For person

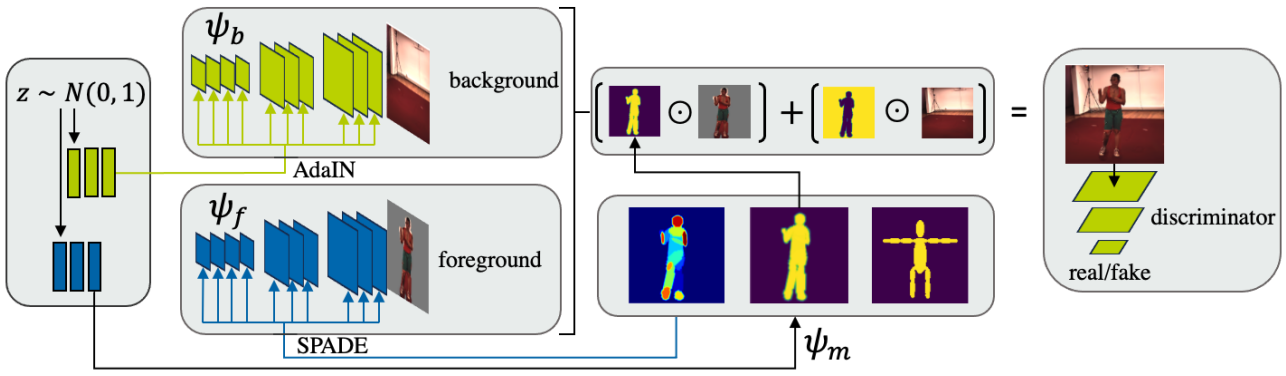



Figure 5.2: Generator architecture: Two separate networks produce fore- and background. The foreground generator outputs the foreground and a corresponding mask. The final image is created via alpha blending fore- and background with the generated mask.

appearance, the latent vector $\mathbf{z}$ gets mapped to an intermediate style code $\mathbf{w}$ similar to [Karras et al., 2018] that is subsequently fed into the foreground generator network. The final image is then obtained via alpha compositing:

$$\mathbf{b} = \psi_{\mathrm{b}}(\mathbf{z});\ \mathbf{m}, \mathbf{p} = \psi_{\mathrm{m}}(\mathbf{z});\ \mathbf{f} = \psi_{\mathrm{f}}(\mathbf{z}, \mathbf{p}) \tag{5.1}$$

$$\mathbf{x}_g = \mathbf{m} \odot \mathbf{f} + (1 - \mathbf{m}) \odot \mathbf{b}, \tag{5.2}$$

where $\odot$ denotes element-wise multiplication.

Once trained, our model can produce realistic images of humans together with pose-landmarks and body part segmentations. This data can then be used to train a fully-supervised model on both landmark detection and human parsing.

### 5.4.1 Generative pathway

#### Template

Similar to previous chapters we make use of a template (see Figure 5.3) where each body part is parameterised by a 2D Gaussian ellipse with spatial centre as well as x- and y-spread defined by the mean and covariance, respectively. In total, our template consists of $K = 16$ connected body parts.

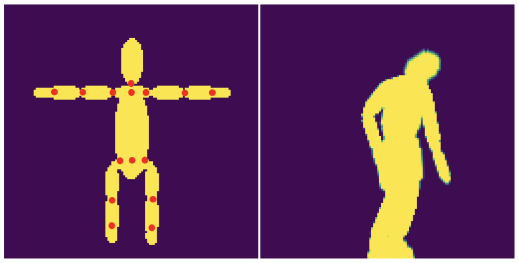


Figure 5.3: Left: Shape template in its standard T-pose configuration and anchor points in red. Right: Template after transformation.

### Template Transformation

To obtain transformed shape templates and masks, a simple MLP (8 fully-connected layers) first maps the input noise vector $\mathbf{z}$ to a set of pose parameters for each of the $K$ body parts defined in the template including rotation ($\mathbf{r}$), scale ($\mathbf{s}$) and translation ($\mathbf{t}$).

$$\mathbf{v}_p = \phi(\mathbf{z}), \mathbf{v}_p \in \mathbb{R}^{66}, \tag{5.3}$$

where $\phi$ is parameterised by an MLP, $\mathbf{v}_p = \{\mathbf{r} \in \mathbb{R}^{K\times 2}, \mathbf{s} \in \mathbb{R}^{K\times 2}, \mathbf{t} \in \mathbb{R}^2\}$. Note that we define translation globally here, *i.e.*, we translate all body parts with a single translation vector. Instead of representing 2D-rotation with a single scalar describing the angle, we instead parameterise by the first column-vector $[\cos(x), \sin(x)]^T$ to ensure a continuous mapping [Zhou et al., 2018].

### Mask

The transformed shape template forms the basis for the binary mask needed for image composition. A straightforward approach would be to simply use the transformed template, *e.g.*, Figure 5.3, as a mask. However, the ellipses, while easy to generate, are only a crude representation lacking details of real human anatomy. We therefore allow the generator to add a limited amount of pixels, which we call the residual $\mathbf{R} \in \mathbb{R}^{h\times w}$ around the boundaries of each part map:

$$\tilde{\mathbf{T}}_k = \mathbf{T}_k + \mathbf{R}_k. \tag{5.4}$$

We then sum individual parts and clip to a maximum of 1 to generate the global image mask:

$$\mathbf{m} = \text{clip}(\sum_k \tilde{\mathbf{T}}_k) \tag{5.5}$$

However, without further restrictions this allows the foreground generator to output a trivial solution with a mask that covers the entire image, *i.e.*, $\mathbf{m} = 1$ everywhere. In order to drive the model towards adding pixels only at the boundaries, we introduce an additional loss that penalises values bigger than 1 outside a margin defined as:

$$\text{margin} = 1 - S(\mathbf{T}_k), \tag{5.6}$$

where $S()$ denotes a scaling operation, *i.e.*, for each part map we increase the size isotropically by a factor $\epsilon$. The resulting map is 1 everywhere except within the enlarged part map. Subsequently, we compute the sum over pixels and template parts of the residual map multiplied with the margin and normalise with the number of non-zero pixels of each part map:

$$\mathcal{L}_{res} = \frac{1}{\sum_{u,v} T_k(u,v)} \sum_k \sum_{u,v \in \Omega} R_k(u,v) \cdot \text{margin}(u,v). \tag{5.7}$$

**Depth-ordering and permutation learning**

A fundamental issue is self occlusion between overlapping transformed 2D part maps. In its existing formulation the model has no way of distinguishing *i.e.*, an arm that is in front of the body facing the camera from one that is behind it. This effectively leads to inaccurate results where the model ignores intersecting template parts and simply paints the final image content onto one of them. Without additional constraints, the model has enough flexibility to do so. We observe this for the aforementioned case where arms are in front or behind the body as illustrated in Figure 5.6. Authors of previous related works have also pointed towards this issue [He et al., 2021]. We aim to solve it via two additions to our model: Firstly, we create

occlusion-aware maps via 2.1-D z-ordering introduced in [Epstein et al., 2022]:

$$\tilde{\mathbf{T}}_k = \mathbf{T}_k \odot \prod_{j=k+1}^{K} (1 - \mathbf{T}_j). \tag{5.8}$$

The index $k$ of each template part can be interpreted as a measure of (inverse) depth. The $k$-th map will be occluded by any of the proceeding maps that cover the same pixels. Intuitively, $(1 - \mathbf{T}_j) \in [0, 1]$ can be interpreted as an opacity map indicating which pixels are partially or fully blocked by other template-parts. However, this approach assumes an ordering between the $K$ template parts and requires the model to freely permute them depending on their position in the depth hierarchy. This in turn depends on the specific pose-configuration. To this end, we propose to let the model learn a permutation matrix $\mathbf{M} \in \mathbb{R}^{k \times k}$. Formally, this matrix contains only 0's and 1's and is *doubly-stochastic*, meaning that all rows and columns sum to 1. A vector multiplied with this matrix contains the same elements, but in a different order. In our case, this means permuting all part maps $\mathbf{T}_k$ by multiplying a broadcasted version of $\mathbf{M}$ with the part-map tensor $\mathbf{T} = \{\mathbf{T}_1, ...\mathbf{T_k}\} \in \mathbb{R}^{K \times h \times w}$

$$\mathbf{T}_{k,perm} = \mathbf{M} \cdot (\mathbf{T}_1, ...\mathbf{T_k}) \tag{5.9}$$

We use the Sinkhorn algorithm as previously applied in the permutation learning literature [Rezatofighi et al., 2018, Mena et al., 2018] in order to learn $\mathbf{M}$. After mapping the latent vector $\mathbf{z}$ to a generic matrix $\mathbf{M}$, the Sinkhorn algorithm achieves an approximation of a doubly-stochastic matrix via iteratively and alternately rescaling rows and columns of $M$. In order to strongly enforce values of either 1 or 0, we additionally penalise the mean entropy along rows and columns of the matrix by

$$\mathcal{L}_e = -\sum_i m_{i,j} \cdot \log(m_{i,j}) - \sum_j m_{i,j} \cdot \log(m_{i,j}). \tag{5.10}$$

**Architectures**

The architecture of fore- and background generators are adopted from StyleGAN [Karras et al., 2019a], where the input noise is mapped to an intermediate style representation that is fed into the convolutional architecture via Adaptive Instance Normalisation (AdaIN) (see section 2.1.3). However, in order to condition the generated content on the human shape template, we replace the scalar parameters for scale and bias with spatially varying input via spatially adaptive denormalisation (SPADE) (see section 2.1.4). We first map the input noise vector $\mathbf{z}$ to $K$ pairs of vectors $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^{K \times d}$, where $d$ is the dimensionality of the intermediate feature space (we choose 512). To encourage better disentanglement between different part maps similarly to [Epstein et al., 2022], we combine these feature vectors with our spatial template part maps through a process coined 'spatial splatting':

$$a_{c,i,j} = \sum_k \mathbf{f}_{c,k} \cdot \mathbf{T}_k, \tag{5.11}$$

where indices $c, i, j$ denote channel and image pixel coordinates, respectively.

**Avoiding all-as-foreground** As noted in [Bielski and Favaro, 2019] and [Yang et al., 2022], training the naive approach with a simple composition leads to trivial solutions where either the fore- or the background contains the entire image. [Bielski and Favaro, 2019] propose to introduce random translations $\Delta_{x,y} \sim U[-\delta, \delta]$ in x- and y-direction in order to force the generator to generate distinct fore- and background objects. However, this rests on the assumption that the position of the foreground object does not correlate with a certain section of the background. In [Bielski and Favaro, 2019], this is the case as the considered datasets are collections of a wide variety of different sources and backgrounds. In our case this assumption does not hold, as we consider many frames from the same video in which both person and background reoccur frequently. When training with random shift, we can observe that there is no perfect separation between fore- and background, as the generator tries to compensate for positional compositions that are not in the original distribution.

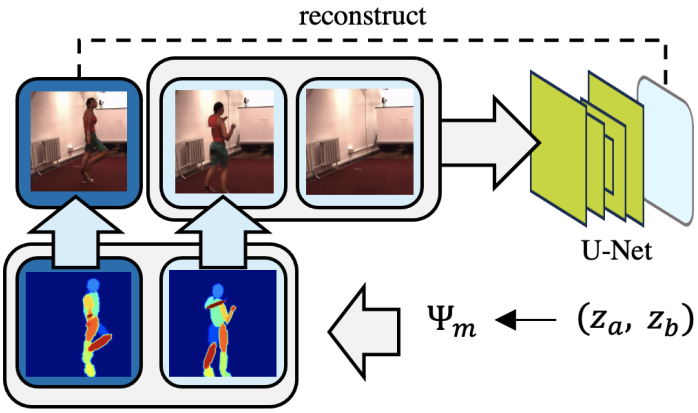



Figure 5.4: In order to drive the model towards more accurate conditioning on all body parts, we include a reconstruction task. An additional U-Net [Ronneberger et al., 2015] needs to reconstruct the original pose given an image with the same person and background but poses swapped.

### Accurate body part conditioning

In its current formulation, the model has the flexibility to generate realistic looking humans without strictly adhering to a corresponding accurate body-part representation via the template as long as the silhouette allows for realistic compositing. This is best illustrated with an example: in a pose configuration where the arms are in front of the torso, the foreground generator can simply paint them as a part of the torso-template section. We illustrate this in Figure 5.6. In order to encourage more accurate conditioning without direct supervision, we add an additional reconstruction objective together with the depth-aware representation illustrated in Figure 5.4. Given two vectors $\mathbf{z_a}$ and $\mathbf{z_b}$ sampled from our prior distribution, we produce template pose representations $\mathbf{p}_a$, $\mathbf{p}_b$. We then generate a composite image from $\mathbf{z_a}$, but swap out the pose conditioning in the foreground generator from $\mathbf{p}_a$ to $\mathbf{p}_b$. Effectively, we create an image of the same person on the same background, but in a different pose configuration.

A U-Net [Ronneberger et al., 2015] type network $\phi$ then receives this image as well as the background image and the original pose template $\mathbf{p}_a$ and is tasked to reconstruct the image created by conditioning on the original pose $\mathbf{p}_a$. This self-supervision pushes the generator to spatially align template and generated foreground, since the reconstruction network relies on accurate information from the template to reconstruct the original image. The objective is similar to to the self-supervision task introduced in the earlier chapters, but with entirely

synthetic data. Since our model is able to separate fore-and background by design, we're not limited to datasets where the background remains constant. Put in equations, we first generate two images $\mathbf{x}$ and $\mathbf{x}_{\text{alt}}$ (the two images displayed in figure 5.4), where the latter is generated by swapping the injected pose representation from $\mathbf{p}_a$ to $\mathbf{p}_b$, keeping both background and style codes $\mathbf{w}$ constant. For brevity, let $G()$ denote the full generation process including alpha compositing, we can write:

$$\mathbf{x} = G(\mathbf{b}, \mathbf{p}_a, \mathbf{w}) \tag{5.12}$$

$$\mathbf{x}_{\text{alt}} = G(\mathbf{b}, \mathbf{p}_b, \mathbf{w}) \tag{5.13}$$

$$\tilde{\mathbf{x}} = \phi(\mathbf{x}_{\text{alt}}, \mathbf{p}_a, \mathbf{b}) \tag{5.14}$$

We train all parts (including the generator network) with a simple $l2$ distance between reconstructed and original image:

$$\mathcal{L}_{recon} = ||\mathbf{x} - \tilde{\mathbf{x}}||_2^2 \tag{5.15}$$

, where $\tilde{\mathbf{x}}$ is the output of the reconstruction network.

### 5.4.2 Training objectives

**Adversarial training** We employ the standard non saturating loss formulated in [Goodfellow et al., 2014b] combined with lazy $R1$-regularisation [Mescheder et al., 2018a] of the discriminator as in [Karras et al., 2019a]:

$$\mathcal{L}_{D,adv} = -\mathbb{E}_{\mathbf{x}_r}[\log D(\mathbf{x}_r)] - \mathbb{E}_{\mathbf{x}_g}[1 - \log D(\mathbf{x}_g)] \tag{5.16}$$

$$\mathcal{L}_{G,adv} = -\mathbb{E}_{\mathbf{x}_g}[\log D(\mathbf{x}_g)] \tag{5.17}$$

**Avoiding all-as-background** As noted earlier, the naive approach to generating fore- and background with separate generators leads to the trivial solution where either of the two con-

tains the entire image and the other remains empty or only makes up a tiny fraction of the composite image. To avoid the background generator from outputting complete images, we propose a simple and effective loss without introducing additional overhead in the form of classifier networks as in [Yang et al., 2022] or additional hyperparameters such as mask size as in [Bielski and Favaro, 2019]. We achieve this by also training the discriminator to classify background images $\mathbf{b}$ as fake while training the generator on the same objective, *i.e.*, aiding the discriminator:

$$\mathcal{L}_{D,b}, \mathcal{L}_{G,b} = -\mathbb{E}_{\mathbf{b}}[1 - \log D(\mathbf{b})]. \tag{5.18}$$

Empirically we observe that training the generator with objectives defined in equations 5.17 and 5.18 leads to improved separation of fore- and background images while avoiding the all-as-background failure mode entirely. An empty background image represents one of the possible local optimums with respect to the two objectives: the composite image (containing the background) should fool the discriminator and therefore has to contain realistic backgrounds, while the pure background image can easily be classified as fake if no foreground object is present.

**Additional losses** In order to drive the model towards binary masks, we employ the following loss similarly to [Bielski and Favaro, 2019]:

$$\mathcal{L}_{bin} = \frac{1}{M} \sum_{i,j} \min(m_{i,j}, 1 - m_{i,j}), \tag{5.19}$$

where M denotes the total number of pixels in the mask.

Additionally, we also employ the boundary loss defined in previous chapters to ensure that transformed template parts stay within the image boundary. We have found that this helps to stabilise early training:

$$\mathcal{L}_{boundary} = \begin{cases} |a_{i,x,y}|, & \text{if } |a_{i,x,y}| > B \\ 0, & \text{otherwise,} \end{cases} \tag{5.20}$$

where $a_{i,x}$ corresponds to the x- and y-component of the $i$-th transformed anchor point.

Furthermore, we regularise the estimated transformation parameters during the first $n$ iterations by adding a decaying $l_2$-norm term for both $\mathbf{r}$ and $\mathbf{s}$ to our overall objective:

$$\mathcal{L}_{reg} = \alpha * (\sum_{K} ||\mathbf{r}_k||_2 + ||1 - \mathbf{r}_k||_2) \tag{5.21}$$

$$\alpha = \min(1/n - 0.01 * \text{iter}) \tag{5.22}$$

**Stage1 - Overall Objective** For stage 1 consisting of training the generator, in summary we optimise the following objective for the generator:

$$\mathcal{L}_G = \mathcal{L}_{G,adv} + \lambda_1 \mathcal{L}_{G,adv,b} + \lambda_2 \mathcal{L}_c \tag{5.23}$$

$$+ \lambda_3 \mathcal{L}_{boundary} + \lambda_4 \mathcal{L}_{reg} \tag{5.24}$$

$$+ \lambda_4 \mathcal{L}_{res} + \lambda_5 \mathcal{L}_{recon}, \tag{5.25}$$

where $\lambda_{1,...5} \in \mathbb{R}$ are hyperparameters for balancing loss terms which we will provide on the supplementary material.

## 5.5 Experiments

After training our factorised generative model, we are able to synthesise large quantities of realistic images and corresponding part-maps and landmarks to train a generic prediction network with full supervision. We provide qualitative and quantitative evaluations in the form of images and the mean pixel-distance between ground truth and predicted landmarks.

**Pose estimation model** We choose a very basic encoder-decoder type network based on a ResNet-backbone [He et al., 2015] similar to [Xiao et al., 2018a]. In line with common practice, we transform the coordinates of the landmarks to Gaussian heatmaps which the model then

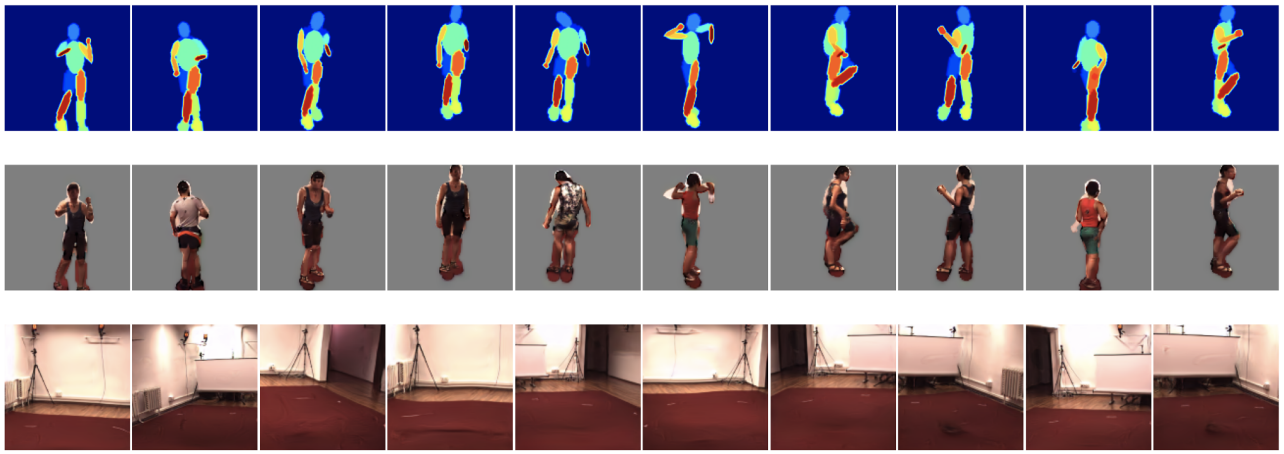


Figure 5.5: Qualitative results from the generative model trained on Human3.6m. Top row: generated body part maps. Middle: generated foreground images. Bottom: generated background images.

learns to regress via a simple mean squared error objective. We employ standard augmentation techniques including random scaling, rotation and horizontal flipping.

### 5.5.1 Datasets

We evaluate our method for 2D pose estimation on a variety of commonly used datasets to demonstrate the generalisability of our approach.

**Human3.6m** As introduced in the previous chapter, Human3.6m is a studio-recorded dataset containing videos captured from different actors with 4 synchronised cameras and corresponding 3D and 2D landmarks defined by optical markers placed on the body. Due to the studio environment, the background variability is very limited, but the dataset serves as a standard benchmark for both 2D and 3D pose estimation.

**YouTube dance dataset** is a collection of 74 single-person dance videos curated by the authors of [Chan et al., 2018], with diverse outfits, backgrounds and challenging poses and varying body types. We focus on qualitative results since there are no ground-truth labels available.

### 5.5.2 Evaluation

We use the Human3.6m dataset and follow the evaluation protocol of previous chapters and use subjects 1, 5, 6, 7, and 8 for training and 9 for testing, resulting in roughly 80000 test frames. Furthermore, in line with previous works, we restrict the dataset to the following activities: *direction, discussion, posing, waiting, greeting, walking*. 16 body landmarks are predicted and compared via the mean euclidean pixel distance normalised by the image dimensions (resulting in percentages). Most landmarks coincide with the anchor points enforcing connectivity of our template, while we add the missing ones (spine and head) in the appropriate locations. It is possible to define an arbitrary amount of points which will get transformed along with each body-part.

Different to other approaches we do not make use of the supplied masks since our goal is to show good fore-and background disentanglement.

### 5.5.3 Results

We summarise quantitative results for our method in Table 5.1. While our method performs considerably worse compared to other weakly and self-supervised methods, it is important to acknowledge that all make use of labels in various degrees either during training or inference and additionally rely on a provided fore- and background separation. Qualitatively, as can be observed in Figure 5.5, our generative model is able to generate background and foreground images as well and to condition itself on accurate part-maps in the absence of any manual supervision. Figure 5.7 displays qualitative results from the prediction model trained on the synthetic data and demonstrates accurate pose landmarks as well as meaningful body part labelling without having to rely on any manual labels during training or inference. Despite this, the boundaries between human and background are not perfectly separated, especially around the feet. The uniform colour of the floor makes it easy for the model to spill a part of the fore- onto the background, since this is harder to detect by the discriminator.

Since no ground truth data is available for the YouTube dataset, we restrict our evaluation

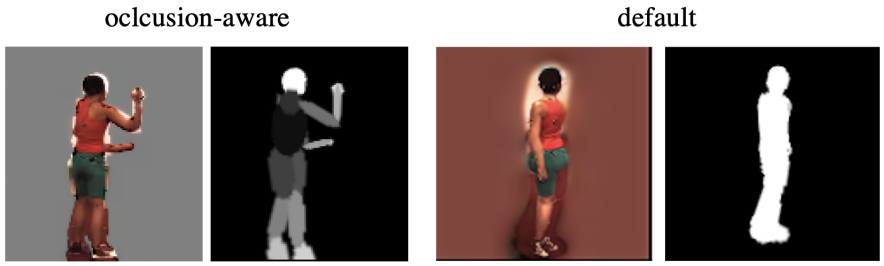



Figure 5.6: Top row: successful example where the left arm is occluded by the body. Bottom row: earlier iteration of the model without occlusion-aware maps. The arm is simply painted onto the image without an accompanying accurate part-map representation

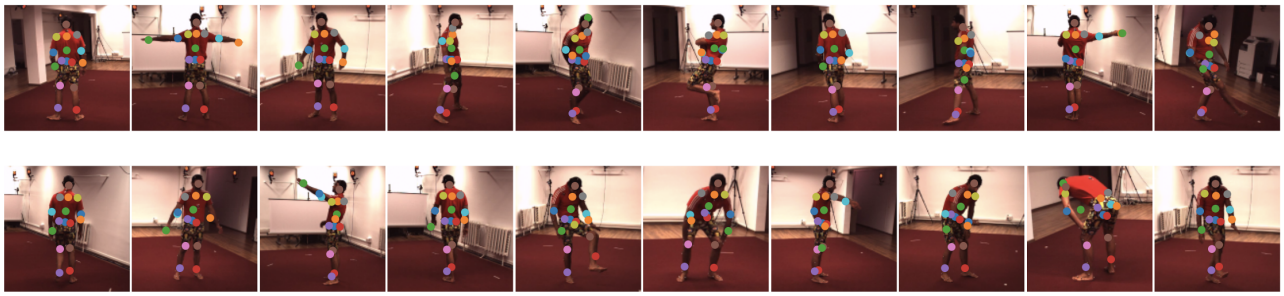


Figure 5.7: Qualitative results from the prediction model trained on generated data from Human3.6m.

to qualitative results shown in Figure 5.8. Due to the very challenging nature of this dataset (complex background scenery, difficult poses and more diverse body proportions) we observe a general degradation in image quality and observe that the model struggles more to properly separate fore- from background.

### 5.5.4 Ablation Studies

Our model consists of a number of components and a combination of different training objectives. We focus on the impact via ablation studies of what we identify as the key novel components introduced in this work, namely:

- The modified adversarial objective for foreground-background-separation
- The handling of self-occlusion via label map permutation learning and depth ordering.

| H36M | all | wait | pose | greet | direct | discuss | walk |
|---|---|---|---|---|---|---|---|
| ***fully supervised baseline*** | | | | | | | |
| [Newell et al., 2016a] | 2.16 | 1.88 | 1.92 | 2.15 | 1.62 | 1.88 | 2.21 |
| ***self-supervised + supervised post-processing*** | | | | | | | |
| [Thewlis et al., 2017a] | 7.51 | 7.54 | 8.56 | 7.26 | 6.47 | 7.93 | 5.40 |
| [Zhang et al., 2018a] | 4.14 | 5.01 | 4.61 | 4.76 | 4.45 | 4.91 | 4.61 |
| [Lorenz et al., 2019a] | 2.79 | – | – | – | – | – | – |
| ***self-supervised (unpaired labels)*** | | | | | | | |
| [Jakab et al., 2019] | 2.73 | 2.66 | 2.27 | 2.73 | 2.35 | 2.35 | 4.00 |
| ***self-supervised (template, no labels)*** | | | | | | | |
| [Schmidtke et al., 2021] | 3.31 | 3.51 | 3.28 | 3.50 | 3.03 | 2.97 | 3.55 |
| Ours | 6.36 | 7.77 | 5.15 | 6.94 | 5.16 | 7.84 | 5.32 |

Table 5.1: Comparison with recent methods for landmark detection on the Simplified Human3.6M dataset; %-MSE normalised by image size is reported on a per action basis.

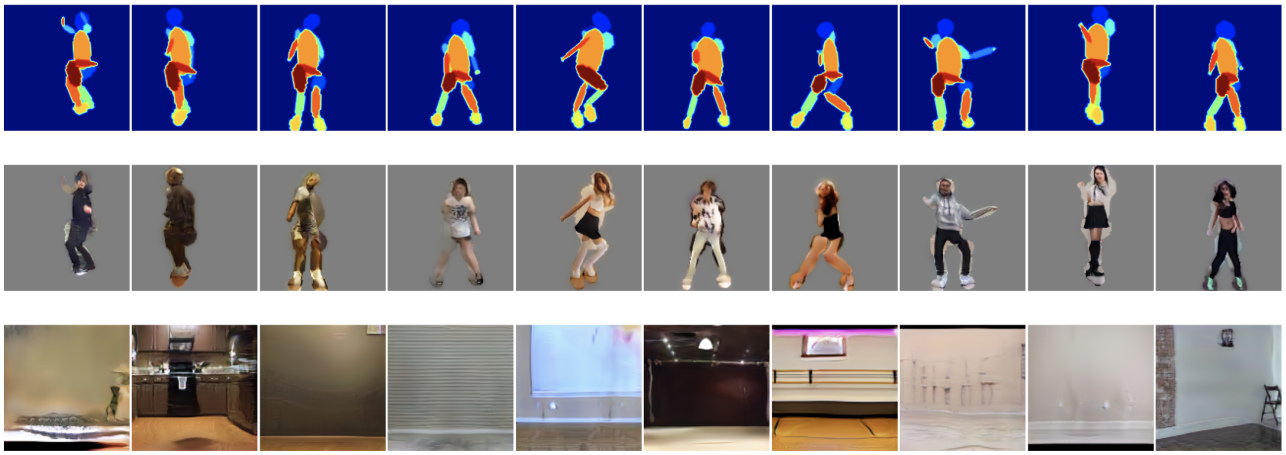


Figure 5.8: Qualitative results from the generative model trained on an in-the-wild setting of dance Youtube videos. Top row: generated body part maps. Middle: generated foreground images. Bottom: generated background images



Figure 5.9: Without modified adversarial training, the model defaults to generating all content via the background. Left: body-part map, middle: foreground, right: background image.

**Modified adversarial objective** We remove Equation 5.17 (training the background generator to aid the discriminator in classifying backgrounds as fake) from the training objective and observe that the model quickly converges to a state where the body-template is scaled down as much as possible and becomes part of the background, while the background generator synthesises the entire image including the foreground. This is demonstrated in Figure 5.9.

**Permutation Learning and Depth Ordering** We train another model without the modifications described in 5.4.1 and display results in Figure 5.6. In pose configurations where the arms are in front or behind of the torso, the model simply paints them as a part of the torso-template part and therefore creates false landmark data.

## 5.6 Additional experiments

The following section introduces additional experiments that either support existing conclusions or explore alternative directions and that were carried out after the first submission of this work.

### 5.6.1 Random shifts

As noted in 5.4.1, random translations $\Delta_{x,y} \sim U[-\delta, \delta]$ of the foreground content and mask can help to avoid 'spillage' of background into the foreground or, in the worst case, a trivial

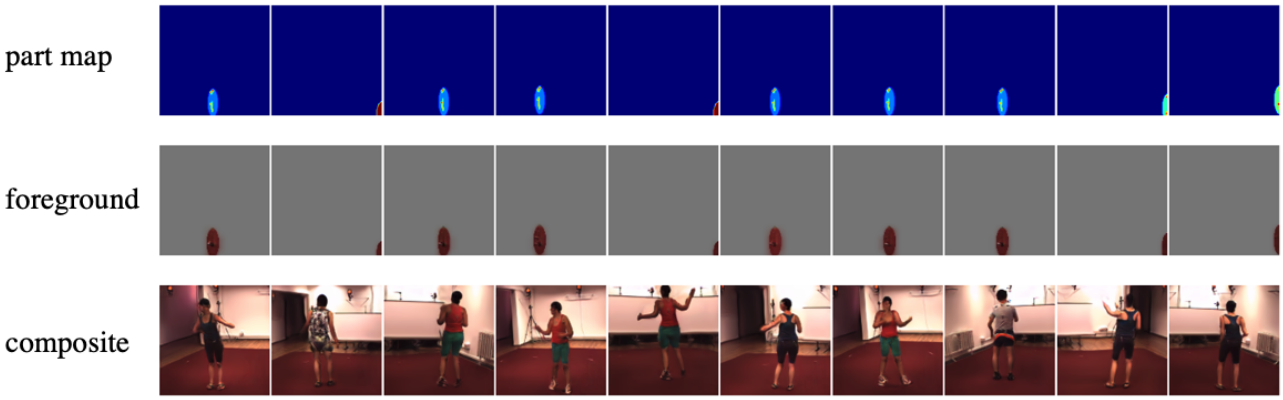



Figure 5.10: Results from the model trained on Human3.6m with random translations, mask size loss and without modified adversarial objective. The foreground completely collapses into a small area and the background generator defaults to generating the entire image.

all-as-foreground solution. The final image becomes:

$$\mathbf{x} = T(\mathbf{m}) \odot T(\mathbf{f}) + (1 - T(\mathbf{m})) \odot \mathbf{b}, \tag{5.26}$$

where $T()$ denotes the translation operation and $\mathbf{m}, \mathbf{f}, \mathbf{b}$ for mask, foreground and background image, respectively. We replaced the modified adversarial objective in 5.17 with a random translation of $\delta = 0.1$ (10% of image height and width). We also included an additional size constraint on the mask similar to [Bielski and Favaro, 2019], incurring cost for the generator when the mask size falls below a certain threshold $\eta$:

$$\mathcal{L}_{\text{size}} = \max(0, \eta - \frac{1}{M} \sum_{x,y} m(x, y)), \tag{5.27}$$

where $M$ denotes the amount of pixels of the mask image.

Qualitative results are summarised in Figure 5.10. With only random shifts, the model quickly converges to a trivial solution where the background generator creates the whole image content, while the foreground is shrunk into a small background area. This further validates the effectiveness of our proposed solution.

### 5.6.2 Disentanglement of pose and appearance

In this experiment we investigate qualitatively how well the generator network can disentangle *pose* in form of the transformed template and *appearance* of a person in the synthesis process. To do so we repeat the process outlined in Section 5.4.1, that is, we generate variation along pose and appearance in isolation by keeping one of the factors constant and varying the other. We can generate images in the following way:

$$\mathbf{x}_i^j = G(\mathbf{b}, \mathbf{p}_j, \mathbf{w}_i), \tag{5.28}$$

where $G()$ again denotes the generation process including alpha compositing and $\mathbf{p}_j$ and $\mathbf{w}_i$ are sampled from different latent codes. Figure 5.11 displays a grid of images where we vary appearance along rows via $\mathbf{w}_i$ and pose along columns via $\mathbf{p}_i$. While the model is able to isolate the two factors in most cases, pose and appearance sometimes remain entangled, as is highlighted via the yellow boxes around some of the resulting images.

### 5.6.3 Estimating pose through inversion

Once all generator networks including $\psi_{\text{fg}}, \psi_{\text{m}}, \psi_{\text{bg}}$ for foreground, background and mask images have been trained to output realistic composite images, in theory it should be possible to estimate pose landmarks from any input image by reconstructing it through the *lens* of our disentangled generative framework. Performing accurate inversion of GANs (as touched upon in 2.2.1) however is a non-trivial task and in fact a sub-field of active research in its own right. Previous works are evaluated along two main criteria: a reconstructed image that remains faithful to the original and an accurate mapping onto the learned manifold or latent-space of the model, enabling downstream tasks such as image-editing. Within the picture of our disentangled generative model we require *both* both criteria to be satisfied: Firstly, the model needs to accurately reconstruct the input image, or at least the person in it, in order to match the pose as closely as possible. Secondly, we need to ensure that the model maintains the disentanglement of background and foreground conditioned on pose, since entanglement of

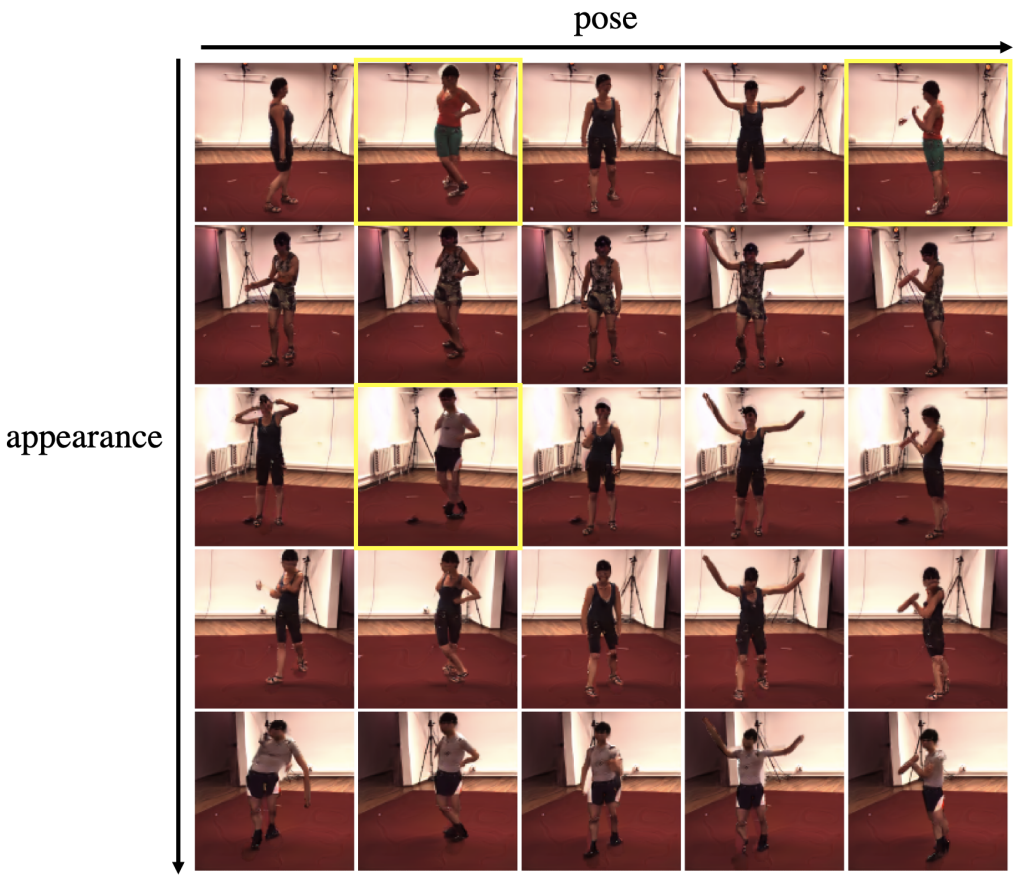



Figure 5.11: Disentanglement results. We vary the appearance in rows via different style codes and the pose in columns through a different transformed shape template. Yellow marked samples denote cases where pose and appearance are still entangled, meaning that a different pose representation leads to a change in appearance despite the style codes being the same.

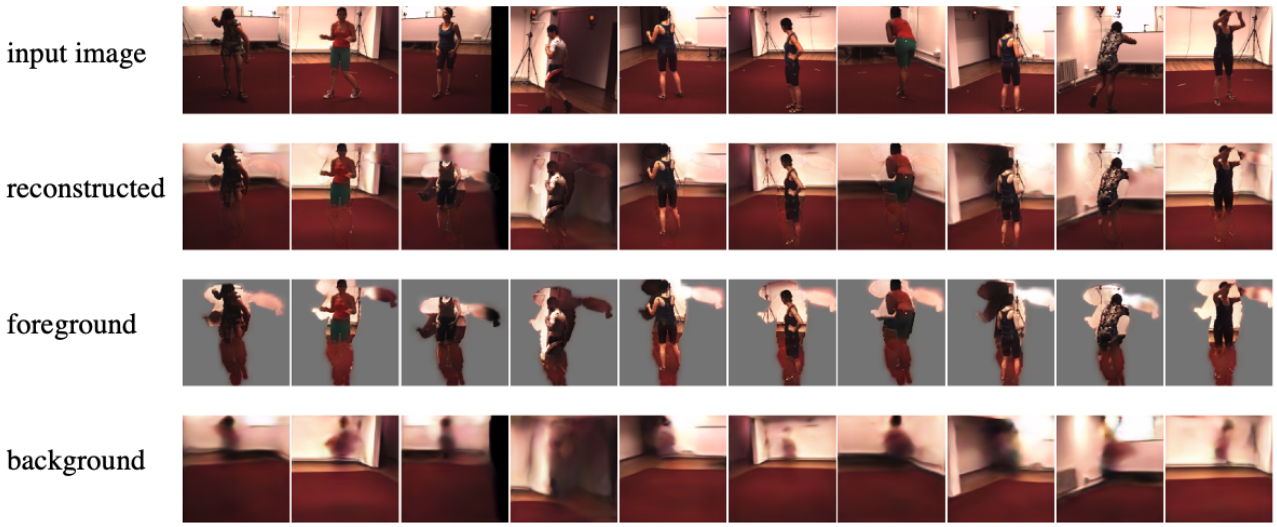



Figure 5.12: Inversion results on Human3.6m with a frozen, pre-trained generator network. Without additional constraints, the disentanglement is broken and the resulting masks and part-maps are no longer meaningful.

these factors will lead to meaningless pose representations even though reconstructions are faithful in terms of pixel-accuracy. Our first inversion experiments in the next section will illustrate this challenge.

### 5.6.4 Frozen generator

#### Encoding into $W^+$-space and direct pose-regression

In our first experiment we follow [Richardson et al., 2020], a popular inversion method for the StyleGAN framework. Figure 5.13 illustrates the principle. We employ feature pyramid networks to estimate style-vectors for background and foreground images as well as transformation parameters $\mathbf{r}, \mathbf{s}, \mathbf{t}$ for the pose template which are then fed into our pre-trained generators $\psi_{\text{fg}}, \psi_{\text{m}}, \psi_{\text{bg}}$ to form a composite image that resembles the input image. The encoders are trained with a simple pixel-wise $l_2$-reconstruction loss between generated and input image:

$$\mathcal{L} = ||\mathbf{x}_{\text{in}} - \mathbf{x}_{\text{recon}}||_2^2 \tag{5.29}$$

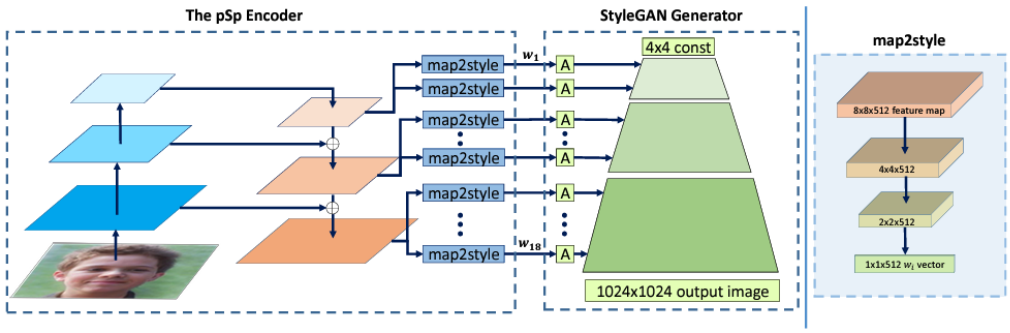



Figure 5.13: Excerpt from [Richardson et al., 2020]. A feature pyramid network [Lin et al., 2016] extracts features from the input image on different scales and maps them to vectors controlling the Adaptive Instance Normalisation in each layer of a trained StyleGAN network.

Qualitative results for Human3.6m are summarised in Figure 5.12. The encoders find style vectors that reconstruct the image regardless of the previously learned disentanglement. This leads to a collapsed pose representation where every input image defaults to the same mask, while both fore- and background elements are simply painted into it. We attribute this to two factors: Firstly, there is no additional constraint on the encoders to force a disentangled representation as in the adversarial training. Secondly, the $W^+$ style-space, while offering good reconstruction quality, introduces too much flexibility for our use-case, to such an extend that the previously learned disentanglement in the generator can simply be 'overwritten'.

### Encoding into the pose latent space

Instead of regressing pose-template transformation parameters directly, we can also map to the latent vector $\mathbf{z}$ before it is fed through the MLP that estimates the transformation parameters $\mathbf{r}, \mathbf{s}$ and $\mathbf{t}$ (see Figure 5.2). However, we observed the same qualitative results as displayed in Figure 5.12.

### Joint training

In this experiment we train both the encoders and the generator network jointly similar to [Han et al., 2021]. A conceptual overview is provided in Figure 5.15. Both generated samples from the generator network as well as reconstructed samples from the encoder are passed

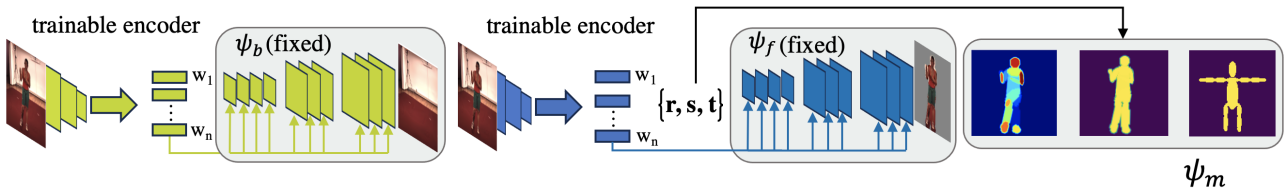



Figure 5.14: We train two encoders, one for the background, and one for the foreground including both the style vectors for appearance and the transformation parameters for the pose template.

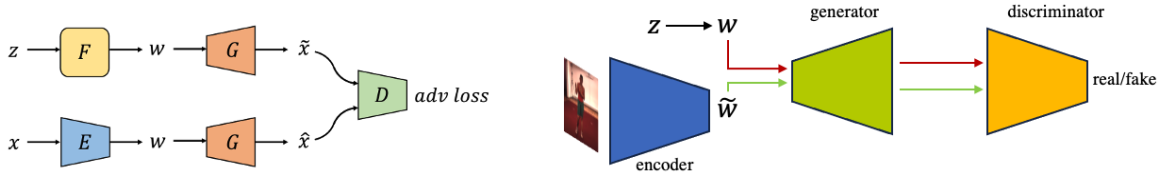



Figure 5.15: Left: excerpt from [Han et al., 2021]. Both encoder and generator are trained adversarially. Right: additional adversarial training: the encoder is trained to match the distribution of generated back- and foreground images.

through the same discriminator network. The adversarial objectives for generator, encoder and discriminator are:

$$\begin{aligned}\mathcal{L}_D = -\mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log(D(\mathbf{x}))] - \lambda_{\text{adv}}\,\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log(1 - D(G(\mathbf{z}))] - \\ (1-\lambda_{\text{adv}})\mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log(1 - D(G(E(\mathbf{x})))]\end{aligned} \tag{5.30}$$

$$\mathcal{L}_G = -\lambda_{\text{adv}}\,\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log(D(G(z))] - (1-\lambda_{\text{adv}})\mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log(D(G(E(\mathbf{x})))] \tag{5.31}$$

$$\mathcal{L}_e = -(1-\lambda_{\text{adv}})\mathbb{E}_{\mathbf{x}\sim p_r(\mathbf{x})}[\log(D(G(E(\mathbf{x})))] \tag{5.32}$$

Additionally, both the generator and encoder are trained on a simple mean squared error reconstruction objective:

$$\mathcal{L}_{\text{recon}} = ||\mathbf{x} - E(\mathbf{x})|\,|_2^2 \tag{5.33}$$

Adversarial objectives for the generator and encoder networks are weighted via the following term:

$$\mathcal{L}_{G,E} = \mathcal{L}_{\text{recon}} + \beta\mathcal{L}_{\text{adv}}, \qquad \beta = \frac{||\nabla_g\mathcal{L}_{\text{recon}}||_2}{||\nabla_g\mathcal{L}_{\text{adv}}||_2}, \tag{5.34}$$

where $\nabla_g$ denotes the mean of gradients of selected layers of foreground and background gen-

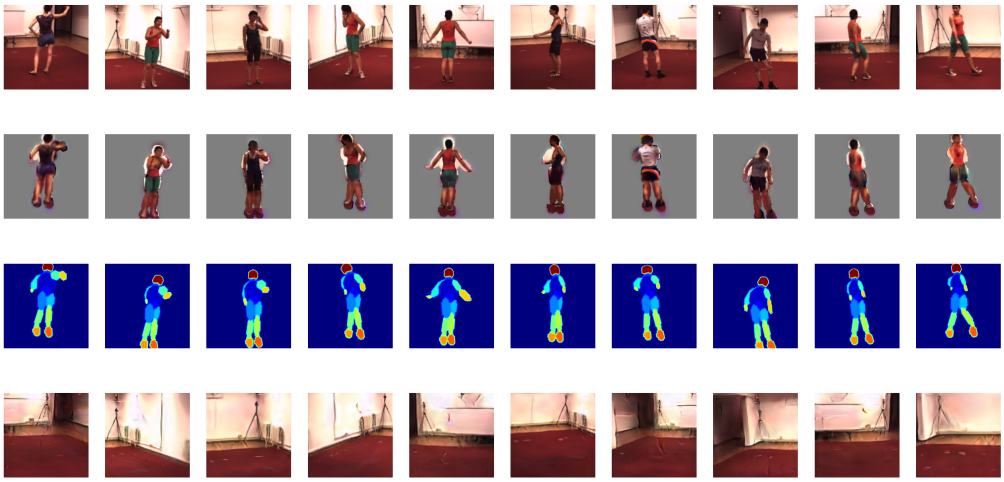


Figure 5.16: Results from the joint model. Top row: input images, second: foreground, third: part-label map, bottom: background.

erators with respect to each of the losses. We choose the last tRGB-mapping layers (see Figure 2.4).

Additionally, in order to enforce disentanglement between fore-and background, we introduce an additional discriminator network that is trained to distinguish generated fore-and background images from reconstructed ones. The losses for encoder and discriminator are analogous to equation 5.34. The left-hand side of Figure 5.15 illustrates the idea. With this additional constraint we ensure that the encoder does not generate foreground content in the background image and vice-versa.

### Results

Qualitative results for Human3.6m are summarised in Figure 5.16. The model successfully disentangles fore- and background from images and is able to extract pose information in the form of the transformed shape template from input images. However, the pose representation remains somewhat inaccurate, partly because we have not included the additional reconstruction objective described in Section 5.4.1.

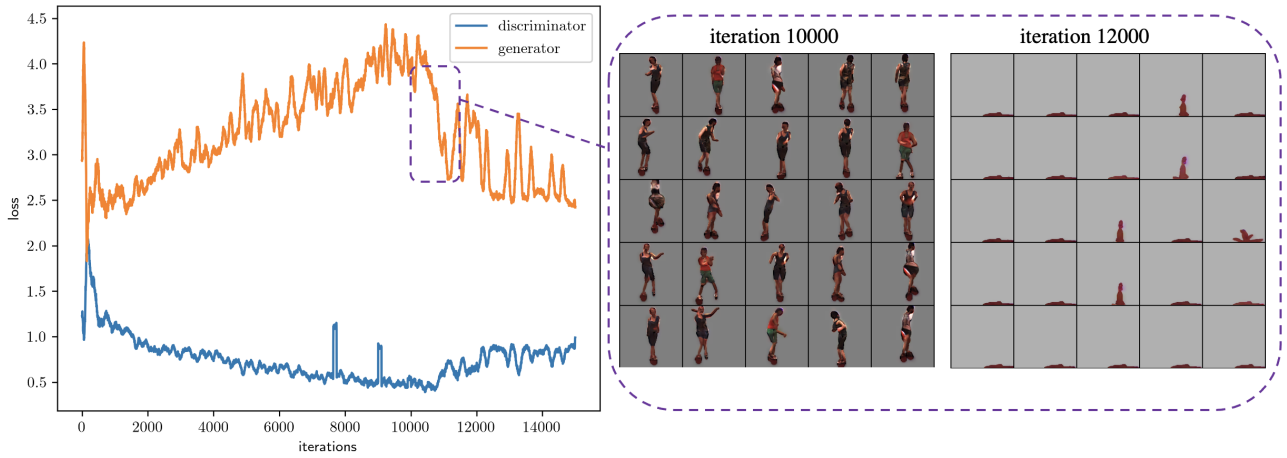



Figure 5.17: Left: loss diagram for generator and discriminator networks. Note that the discriminator loss is not exactly mirroring the generator loss because both are the result of a sum of individual components. Right: generated foreground images at different iterations.

### 5.6.5 Improving segmentation

As noted earlier, the separation between fore-and background is not perfect and becomes more pronounced in more challenging datasets as visualised by Figure 5.8. Since fore-and background generator share the same latent code, both can and will synthesise parts of the image that we wish to disentangle without the use of additional objectives. In this section we provide experiments for several ideas to tackle this challenge.

#### Separate latent codes

One obvious experiment is to generate fore- and background with separate latent codes $\mathbf{z}_{\text{fg}}$ and $\mathbf{z}_{\text{bg}}$. This however carries the implicit assumption that foreground object and background are uncorrelated, which is certainly not the case for diverse enough datasets including our dance dataset collected from YouTube or even Human3.6m, where despite the studio setup, the exact appearance of the person depends on lighting which in turn depends on the positioning in the room. We can clearly observe in Figure 5.17 that until around 10000 training iterations, the classification loss for the discriminator decreases while the cost for the generator grows analogously. A sudden change occurs between iteration 10000 and 12000, where the foreground fully collapses into a part of the background. This provides evidence that with separate latent

codes for fore-and background, the foreground generator misses essential information about the background image in order to create a realistic composition, which the discriminator can exploit during the adversarial game. In fact we observe that the foreground representation almost instantly collapses when training on our YouTube dataset, since there is a strong correlation between subject identity (clothes and overall appearance) and background beyond more subtle changes in lighting. However, the generated foreground images displayed in Figure 5.17 clearly show less spillage of background elements into the foreground, as the foreground generator has no knowledge about the specific latent variable that produced the background image.

**Shrinking the mask**

If the foreground and the corresponding mask are spilling into the background, we can attempt to shrink the mask to the boundaries between background and foreground object. These boundaries, however, are not explicitly known, which makes the task more difficult. In essence, we need to shrink the mask, but only up until the boundary. The naive approach would be to make an estimate of the mean ratio between the sizes (in terms of number of pixels) of shrunk and original masks, *e.g.* 0.8, and design a fitting morphological erosion operation that we apply to every mask. But this simple heuristic does not account for the large variances in individual shapes and would lead to inaccurate results.

An alternative approach would be to ask: *How much can we shrink the mask without hurting the realism of the composite image?* Luckily, we have a mechanism for this in the framework of our generative model: we can expose the composition of background and foreground with the shrunk mask to the scoring of the discriminator network. Any redundant pixels that are already modelled in the background image should have no impact on the discriminator score when being removed from the foreground image. We define the mask shrinking via an image-to-image mapping $\Lambda : \mathbb{R}^{h \times w} \to \mathbb{R}^{h \times w}$, where $h$ and $w$ denote height and width of the mask image. We parameterise this mapping with a neural network and choose the U-net architecture because we need to it helps to facilitate propagation of dense pixel-information to the output. We restrict the output to lie in the range of $[0, 1]$ and obtain the final mask by substracting

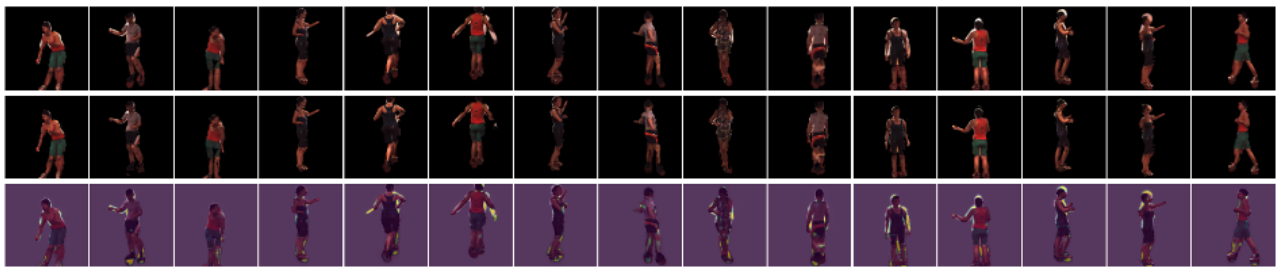


Figure 5.18: Qualitative results of our mask refinement model trained on Human3.6m. Top row: generated foreground images multiplied with the original mask, middle row: foreground images multiplied with the refined mask. Bottom row: foreground images overlaid with parts of the mask that are removed in green.

$\Lambda(\mathbf{m})$ from the original mask $\mathbf{m}$ as well as clipping the result to positive numbers:

$$\tilde{\mathbf{m}} = \mathbf{m} - \Lambda(\mathbf{m}) \tag{5.35}$$

For optimisation, we introduce to objectives. Firstly, we task the network to minimise the size ratio (in our case the sum over pixel values since our masks are approximately binary) between shrunk and original mask. The second objective corresponds to the generator loss in the GAN adversarial game, where we try to maximise the cost of the discriminator.

$$\mathcal{L}_{\text{shrink}} = \frac{\sum_{i,j} \Lambda(\mathbf{m})(i,j)}{\sum_{i,j} m(i,j)} \tag{5.36}$$

$$\mathcal{L}_{adv} = -\mathbb{E}_{\mathbf{z}\sim q(\mathbf{z})}[\log(D(C(\mathbf{b}, \tilde{\mathbf{m}}, \mathbf{f}))] \tag{5.37}$$

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda \mathcal{L}_{\text{shrink}} \tag{5.38}$$

, where $C(\mathbf{b}, \tilde{\mathbf{m}}, \mathbf{f}))$ denotes the alpha-composition operation with background $\mathbf{b}$, foreground $\mathbf{f}$ and shrunk mask $\tilde{\mathbf{m}}$ and $\lambda$ is a hyperparameter for weighting objectives.

Qualitative results are summarised in Figure 5.18. We observe that our mask refinement network is capable of identifying and removing redundant background elements, leading to a more accurate segmentation. However, we sometimes also observe small holes in the segmentation where the model incorrectly removes foreground pixels from the mask. A rather severe case of this can be seen in column number 8 from the left in Figure 5.18, where significant

parts of both legs as well as the left arm are incorrectly removed. We attribute this to the two fundamentally contrasting objectives, *i.e* maximising the total mask area removed versus retaining realism in the composite image.

However, the corrections to the segmentation mask do not mitigate wrong pose landmarks due to the transformation of the shape template. To illustrate, let us consider the sixth column in Figure 5.18, large parts of what is supposed to represent the right arm are removed because they contain background pixels, while the underlying pose representation remains the same, meaning that elbow and wrist landmarks are incorrectly placed. Therefore, in the next experiment we try to incorporate the signal from the mask refinement network into the learning objective of the generator. Our qualitative results suggest that the mask refinement network can effectively identity background regions. Our hypothesis is that simultaneously training the generator to maximise the loss of the refinement network leads to a lower proportion of generated background regions and ultimately better correspondence between the generated person and pose landmarks. From the point of view of optimisation, we include the inverted cost function of the refinement network into the overall cost function of the generator:

$$\mathcal{L}_{\text{refine}} = 1 - \mathcal{L}_{\text{shrink}}, \tag{5.39}$$

where $\mathcal{L}_{\text{shrink}}$ is the ratio of refined to original mask as detailed in equation 5.38. In other words, we train the generator to maximise this ratio to 1 such that the refined mask is equal to the original, *i.e.* the refinement network is unable to to identify background pixels that can be removed from the foreground image without sacrificing realism.

Figure 5.19 shows a qualitative comparison between the original model and the model trained with the objective described in the previous paragraph after 15000 training iterations. Overall we observe less spillage of background elements into the foreground. However, the loss defined in equation 5.39 starts to increase again as displayed in Figure 5.20, while we observe a slight decrease in the adversarial cost of the generator (*i.e.* the loss associated with fooling the discriminator into classifying generated samples as real).

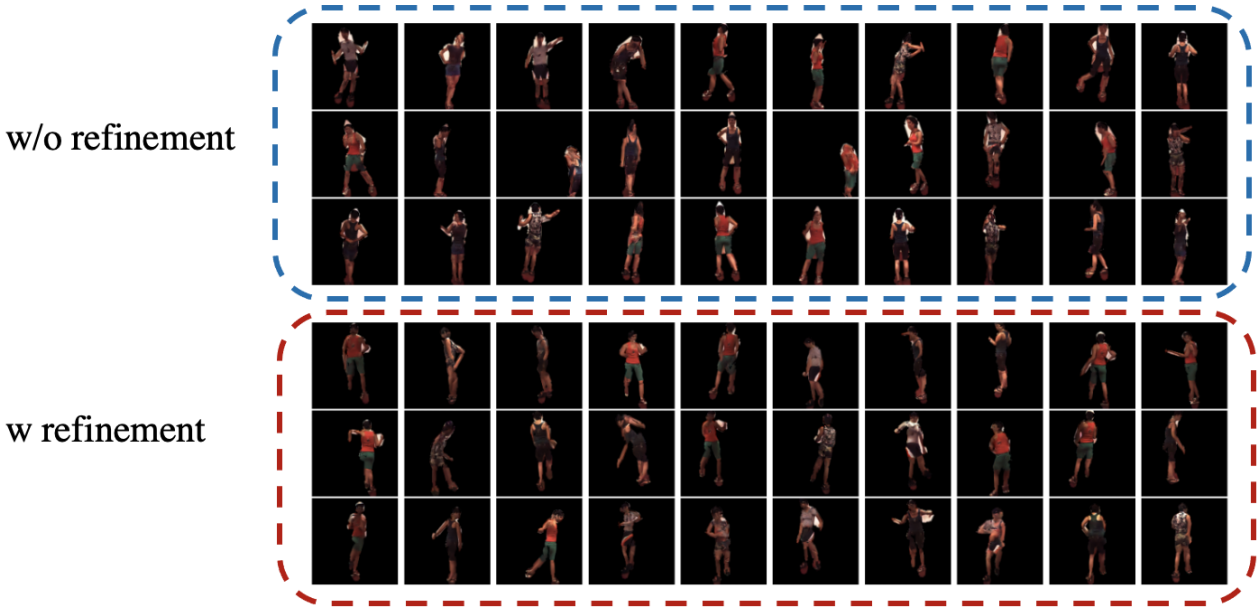



Figure 5.19: Generated foreground image multiplied with the mask for two different models at an intermediate training stage of 15000 iterations. Top: original model, bottom: model trained with cost function defined in equation5.39. The latter generates less background elements as part of the foreground mask

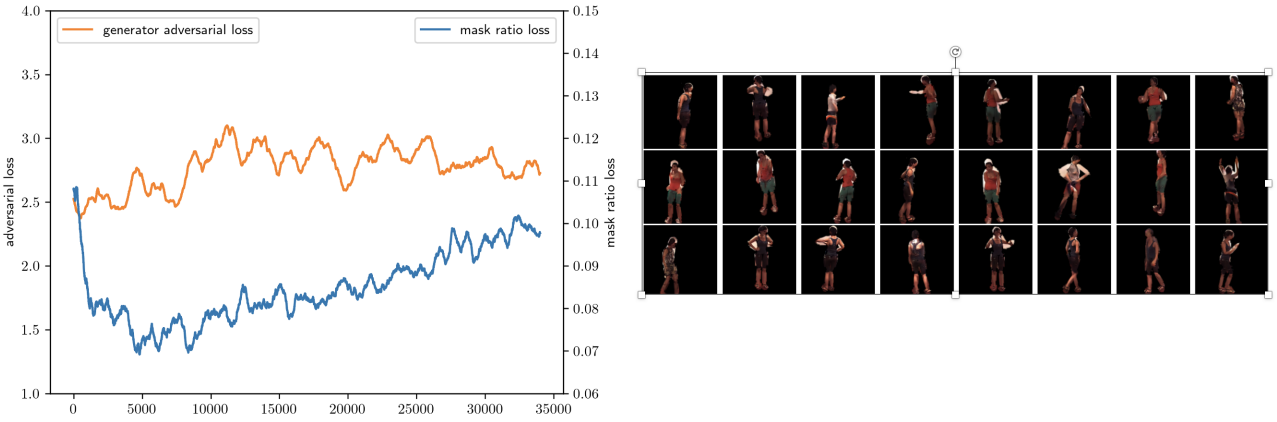



Figure 5.20: Right: two overlaid losses during training: adversarial loss of the generator and mask refinement loss (see equation 5.39). Left: Qualitative results in the form of generated foreground image multiplied with the mask at the final stage of training after 59000 iterations.

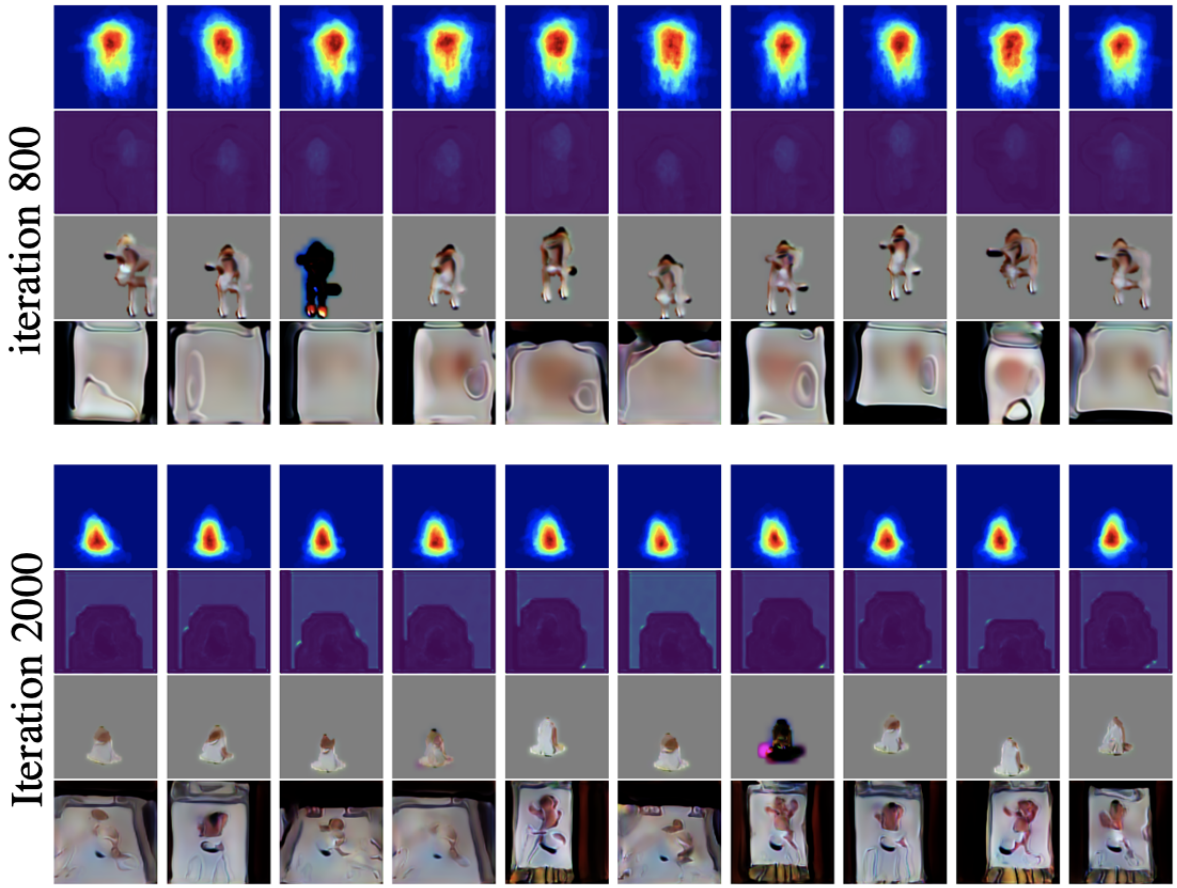



Figure 5.21: Qualitative results of our model trained on the infant dataset described in section 3.3 of the previous chapter during early iterations of training. Top row: pose representation, second row: template residual, third row: generated foreground image, bottom row: generated background image. The model defaults to a all-as-background solution and learns to transform the template into a very small area by overlapping parts via rotation.

### 5.6.6 Results on infants

To test generalisability to different body proportions, we also test the method introduced in this chapter on our dataset of videos from newborn infants introduced in section3.3 of the previous chapter. Figure 5.21 shows generated fore- and background images as well as transformed pose templates through various stages of early training. Despite our introduced mechanisms (see section 5.4.2 for details) the model defaults to collapsing the template and generating all relevant image content in the background.

## 5.7 Summary

We proposed a model that generates realistic images of humans, factored into foreground, background as well as an interpretable landmark and body part representation without relying on labels. The synthesised data can be used to train fully-supervised models on human pose estimation or body-part segmentation. We also found and investigated current limitations, most prominently imperfect fore- and background separation which becomes more pronounced on the more challenging YouTube dataset with more complicated and varying poses and backgrounds. Additionally, we observed a failure of the model to converge on certain datasets such as on infants. In further experiments we investigated and qualitatively validated a novel strategy to improve the segmentation, but found that despite encouraging qualitative results in the early and intermediate stages of training, the model is pushed towards incorporating background pixels into the foreground. We also showed results for an alternative pose estimation approach based on the inversion of the generative model.

## 5.8 Appendix

:

### 5.8.1 Training and hyperparameters

We train our model on 4 Nvidia A100 GPUs in the distributed setting and the following hyperparameters:

- **Batch size:** (single GPU) batch size: 32, effective: 4 x 32 = 128
- **Optimiser:** We use Adam [Kingma and Ba, 2014] with learning rate 0.0001 and $(\beta_1, \beta_2) = (0, 0.9)$. Setting $\beta_1 = 0$ was crucial for stable training
- **Regularisation:** We employ lazy $R1$-regularisation of the discriminator similar to [Karras et al., 2019a] every 8-th iteration.

- **Loss weigthing** We use $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1, \lambda_4 = 5, \lambda_5 = 2$ and $\lambda_6 = 20$.

#### Residual loss

For the residual loss described in Equation 7 in the main paper, we linearly increase the contribution to the overall loss from 0 to 1 over 5000 iterations. We found that this helped to enforce the residual maps to stay within the margin without driving the model to output empty maps immediately.

### 5.8.2 Architecture

#### Foreground generator

A detailed overview of the foreground generator architecture is displayed in 5.22. Following [Epstein et al., 2022] we first map the gaussian prior to *style* and *structure* vectors. Different from [Karras et al., 2019a], we start from a resolution of 16x16 instead of 4x4.

#### Discriminator and background generator

Both discriminator and the background generator have the same architecture (with fewer overall feature channels and decreased resolution) as described in [Karras et al., 2019a].

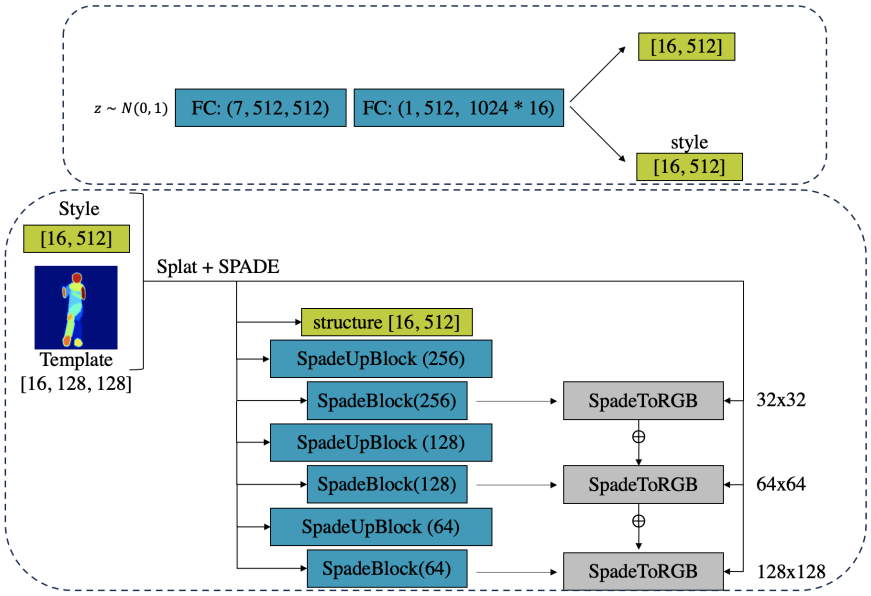



Figure 5.22: Detailed architecture of the foreground generator. The style vector and and template are combined via the 'splatting' operation described in equation 11 in the main paper. Squared brackets denote dimensions of tensors, curly brackets describe the number output channels of the convolutional layers. For the FC-layers, the numbers denote number of layers, input channels and output channels, respectively.

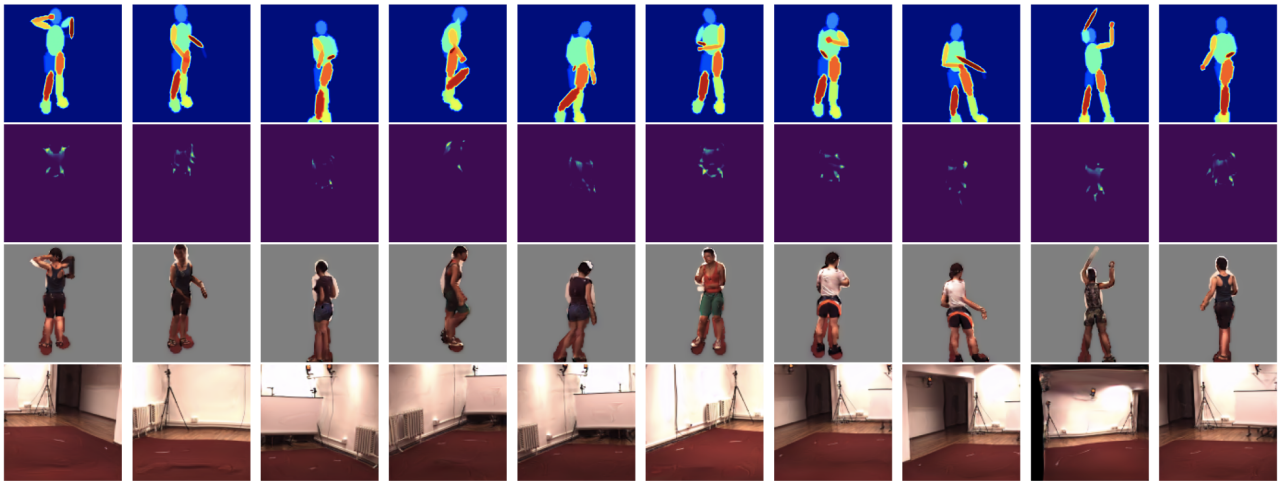


Figure 5.23: Additional generative results of the model trained on Human3.6m. Top row: generated body-part maps. Second row: Sum of the residual maps added to the mask. Third row: generated foreground images. Fourth row: generated background images

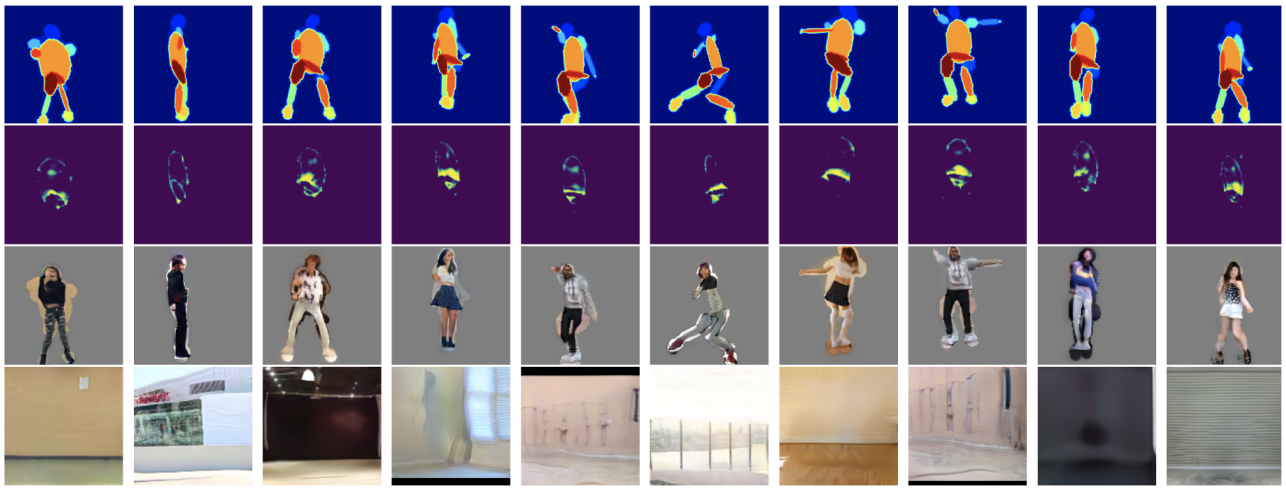


Figure 5.24: Additional generative results of the model trained on the YouTube-Dance dataset. Top row: generated body-part maps. Second row: Sum of the residual maps added to the mask. Third row: generated foreground images. Fourth row: generated background images

# Chapter 6

# Discussion & Conclusion

## 6.1 Discussion

As outlined in the introduction, the main objective of this thesis was to develop self-supervision for label-efficient human movement analysis from image and video data. In a wider context, we tried to approach the question of how a computer-vision model can learn the *decompositionality of the human body into connected parts* without the need for landmark annotations. Throughout the thesis we explored different options including conditional image translation, generative modelling as well as 2D and 3D connected shape templates as a powerful prior.

**In the third chapter** we introduced a 2d shape template consisting of Gaussian ellipses connected via anchor-points. Pose was learned as a factor of variation in videos with a static background in a conditional image translation or reconstruction task. We were able to show substantial improvement over previous work by eliminating the need for unpaired pose labels, which become a considerable weak-point when there are significant domain shifts, as we demonstrated in quantitative experiments on infant videos. However, we also found limitations of our method that sparked ideas leading to alternative or extended approaches presented in the subsequent chapters. The perhaps biggest drawback is the requirement for a constant background, which severely restricts the applicability of the method in real world scenarios. On the other

hand, the training objective is prone to errors when the variety in pose between pairs of images is not high enough. We found this to be especially apparent when comparing estimated landmarks for hands and feet on the infant dataset. Feet markers are consistently estimated more accurately than hands, which can be observed qualitatively in images (3.6) as well as quantitatively in the computed metrics broken down by location (3.3). When watching videos of these infants, it becomes clear that there is overall a lot more variety in leg movement compared to the arms, leading to a better self-supervision training signal for the lower part of the body.

**The fourth chapter** further extends the self-supervision concept from the first chapter with a 3d shape template and explores different differentiable rendering techniques including mesh-based representations in order to train the template-based pose representation on an image reconstruction task in an end-to-end manner. For the mesh-based approach, we chose a SMPL body model Loper et al. [2015b] adjusted to infants Hesse et al. [2018b] but were not successful in deriving meaningful pose representations. Our hypothesis is that this is due to two factors. Firstly, the body in SMPL is described by 72 parameters for pose and 20 for shape, which results in an overall larger parameter space compared to simpler body models, making regression more difficult. Secondly, the rendering process might not produce suitable gradients for end-to-end learning, as the soft rasterisation procedure is only an approximate remedy for the discontinuities hindering backpropagation in the original process.

We found that volume ray casting provides the best results and showed qualitative evidence that 3d shapes and pose can be learned from 2d images without manual supervision and purely driven by an image reconstruction loss. Despite this, we also encountered problems including depth ambiguity and errors due to pairs of training images not including enough variety between poses as discussed in the previous paragraph. A straightforward improvement would be weak supervision in the form a small labelled dataset, pre-training on synthetic data or input images from multiple views.

**The fifth chapter** introduced generative modelling as a means to learn a pose representation irrespective of the background in an effort to remove the biggest limitation of the methods introduced in the first and second chapters. By learning a generative model that can compose an image via disentangled factors including background, human appearance as well as pose we explored training a pose detector model on synthesised data as well as estimating pose directly via image reconstruction through inversion. We discovered a new effective way for fore- and background disentanglement in GAN-models in general as well a possible path towards fully unsupervised, anatomically meaningful part-segmentation of the human body. However, we also identified some remaining challenges: the separation between fore- and background is not perfect, and often elements of the background are modelled as part of the foreground, which results in inaccurate global and individual part masks. Moreover, generating a synthetic pose landmark dataset and subsequent training with a standard, fully-supervised pose detector yielded worse results compared to related previous methods. However, it is important to note that all methods we compared to either need some degree of supervision to map to final landmarks or are constrained in significant ways, *i.e.* requiring a constant background.

The exact cause for the worse performance remains an open question, but we can reason about at least two possible sources of error: Firstly, the aforementioned imperfect alignment between synthetic pose landmarks and image due to sub-optimal fore- and background separation propagates into the training of the pose detector. Secondly, GANs are notorious for mode-collapse and there exists a trade-off between sample fidelity and diversity Karras et al. [2018]. If the distribution of generated images and corresponding poses is not diverse enough, it could negatively affect the pose detector's generalisation capabilities when trained on such data.

The complete failure of the model on the infant dataset sparks additional questions that need to be investigated in future works. Our hypothesis is that the position of the infant is highly correlated with an element of the background, namely the cot. In comparison, there is less correlation between human and background in both the Human 3.6m as well as YouTube datasets, preventing the discriminator from exploiting placement information. We tried giving the mask

and foreground generators access to intermediate, low-resolution feature maps of the generated background to regress global translation parameters from, but this led to entanglement of back- and foreground pixels. This leads us to suspect that some degree of positional independence between foreground object and background as assumed in Bielski and Favaro [2019] and discussed in Section 5.4.2 of Chapter 2 is necessary for the model to perform well.

In a wider context and independent of architectural choices and current limitations, our results demonstrate that generative models can learn meaningful hidden structure including, but not limited to, pose and part-segmentations from unlabelled image data, given the right priors and training objectives.

## 6.2 Outlook

Generally speaking, we strongly believe that the approach presented in Chapter 3, *i.e.* learning human pose through discovering the inherent compositionality of the human body from generating images is a very promising direction of research. In the following paragraphs we will discuss specific possible directions in more detail.

**Improved disentanglement of fore- and background** As evidenced by our experiments in Chapter 3 (5.6.5), we believe that it is possible to achieve a better separation of fore- and background via our refinement approach. The results displayed in figure 5.19 are encouraging, but the generator still resorts to generating background elements during a later stage in training. The accompanying decrease in the adversarial cost for the generator is an indicator that with the current architecture, a more accurate separation is only possible at the expense of image realism. This in turn leads us to believe that the problem discussed in section 5.4.1 persists: The Gaussian ellipses that form different body parts are not suited to represent fine-grained details such as a pixel-accurate border between human and background. Despite our introduction of additional pixels that can be added to each part mask, the balancing of different cost functions to keep a representational bottleneck remains challenging. For the mask refinement network, we

did not explore different architecture choices and opted for a simple U-Net Ronneberger et al. [2015], but different architectures, especially the cross-attention mechanism utilised to great success in the latent-diffusion model Rombach et al. [2021], might bring further improvement. In fact, one might also consider swapping the entire mechanism by which we feed the pose representation into the network, namely SPADE Park et al. [2019b] in combination with Epstein et al. [2022], with the cross-attention mechanism.

**Improved disentanglement of pose and appearance** The results in figure 5.11 show that our generative model introduced in chapter 3 is capable of disentangling pose and appearance, albeit not perfectly. Strong correlations between unique movements and the actors performing them sometimes carry over into our model. Given that we have explicit control over the transformation of the template, it might be an option to include equivariance constraints as explored in Lorenz et al. [2019a] to improve the disentanglement.

**Weak supervision** The set goal throughout this thesis was to push the boundaries of what is achievable via the minimum amount of annotations during training. However, quantitative results of the method presented in chapter 3 are not convincing enough yet. Adding some form of weak supervision, *e.g.* landmark annotations for a limited amount of images, synthetic data or multiple views for estimation in 3D could produce quantitative results that are stronger than existing work. Text-guided diffusion models have taken the field of image generation by storm, partly because they are being trained on pairs of images and text descriptions in the order of billions. A plethora of novel research utilises models such as latent diffusion Rombach et al. [2021] trained on these datasets of unprecedented scale as a backbone for downstream applications, demonstrating that the model learned very strong priors. It would be interesting to investigate to what degree such a model has learned pose and appearance as separate factors and how such a disentanglement can be further facilitated.

## 6.3 Conclusion

In this thesis we aimed to push the current boundaries of self-supervised learning in the context of human pose estimation from images. More specifically, we asked what is the absolute minimum amount of supervision required to extract meaningful pose representations. It is not an overstatement to say that we went further than any previous work in answering that question, and demonstrated feasibility in some contexts and limitations to overcome in others. We are certain that future work will address the remaining challenges and move us closer towards not just fully self-supervised and accurate human pose estimation and tracking, but truly intelligent systems that can learn from vast amounts of data without the need for human supervision.

# Chapter 7

# Appendix

## 7.1 Computer vision, machine learning & deep learning

We precede the introduction into relevant concepts by giving broad definitions of different scientific disciplines.

**Computer vision** is a scientific discipline concerned with the recognition and interpretation of image data in any form, for example coming from any type of camera, scanners, microscopes or medical imaging devices. There are many different tasks of interest including the recognition and classification of objects, tracking, segmentation, localisation and mapping and generating or altering image content.

**Machine learning** , in its broadest definition, describes the use of algorithms running on computing devices that can be trained to act on an input without explicit instruction Koza et al. [1996]. Most commonly, algorithms are categorised depending on the nature of the learning signal they are trained with:

- *Supervised Learning*: Given data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and desired outcomes (labels) $\{\mathbf{y}_1, ..., \mathbf{y}_n\}$ coming from an external source, the goal is to train a model to map data to the desired

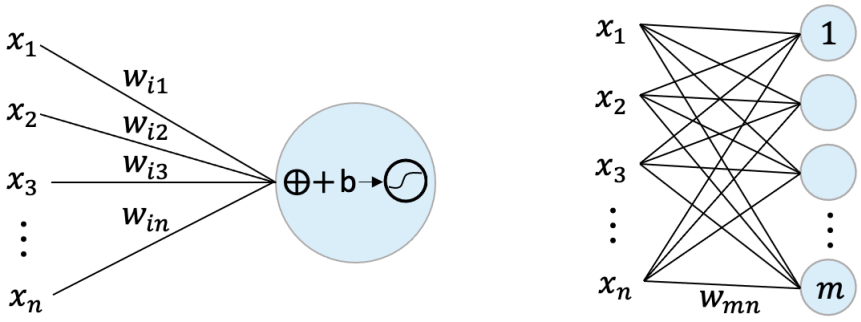



Figure 7.1: Left: schematic display of an artificial neuron. Right: several artificial neurons stacked into a layer.

outcome, most commonly through minimising an associated cost-function $\mathcal{C}$. Often, finding the particular cost function that leads to the desired outcome is a non-trivial task in its own.

- *Unsupervised Learning*: In contrast to supervised learning, the model has no access to external labels during the learning process. One prominent strategy is *Self-Supervision*, in which labels can be found that are part of the data itself along with a proxy supervised task, that, once learned by the model, generalises to or aids in finding the final desired outcome.

**Deep learning** is a sub-field of machine learning concerned with the use of a particular class of algorithm or model, the Artificial Neural Network. The term 'deep' herein refers to several layers of neurons stacked behind each other, such that the output of one layer gets fed to the input of another until the final output has been reached. As the name suggests, Artificial Neural Networks are inspired by biology in the form of connectivity patterns found between neurons, the cells that are responsible for the electrical signal propagation responsible for brain and nervous functions in animals.

### 7.1.1 Artificial neural networks

#### The artificial neuron

The artificial Neuron (see Figure 7.1) forms the basis of the artificial network. Given an input vector $\mathbf{x} \in \mathbb{R}^n$, the $i$-th neuron computes a weighted sum of each of the $n$ components, adds a bias $b$ and passes the output to a non-linear activation function $\phi$:

$$o_i = \phi(\sum^{n} w_{in} x_n + b) \tag{7.1}$$

As demonstrated in Krizhevsky et al. [2012], the choice of non-linear activation function is crucial for faster convergence. The so called Rectified Linear Unit (ReLU in short) Jarrett et al. [2009] has become the standard choice:

$$\phi(x) = \max(x, 0) \tag{7.2}$$

#### A layer of neurons

Several neurons can be combined into a layer (see Figure 7.1). The output computation can be conveniently expressed in vectorised form by a simple multiplication between the weight matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ and the input vector $\mathbf{x}$:

$$o = \phi(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{7.3}$$

#### Neural networks

The layer described in the previous section forms the basic building block for *Neural Networks*, that is, several layers with varying numbers of neurons stacked on top of each other. The final output is the result of feeding the output of layer $l_i$ to the next layer $l_{i+1}$. Put more formally, the final output $\mathbf{o}_{\text{final}}$ of a neural network with $n$ layers is the result of a concatenation of neuron

layer computations:

$$\mathbf{o}_{\text{final}} = l_n(l_{n-1}...(l_0(\mathbf{x}))) \tag{7.4}$$

The connectivity pattern between neurons and input as well as the sequential order of computation described here present only one of many possible configurations that depend on the use-case and nature of the input data.

### 7.1.2 Convolutional neural networks: A strong prior for images

#### Convolutional filters in digital image processing

Mathematically, a two-dimensional image can be described as a function $f : \Omega \subset \mathbb{N}^2 \to \mathbb{R}$ that maps each location within a two-dimensional grid to a scalar value. Image filtering is a standard technique in digital image processing that allows to enhance, suppress or alter certain spatial frequencies in the image. It is most commonly applied to reduce high-frequency noise (low-pass filtering) or enhancing certain details including edges and boundaries (high-pass filtering). Viewed in the spatial domain, image filtering can be described mathematically by convolving a two-dimensional kernel function $k : \Lambda \subset \mathbb{R}^2 \to \mathbb{R}$ with the image:

$$\tilde{f} = f * k \tag{7.5}$$

The two-dimensional convolution operation in its discrete form can be expressed as:

$$\hat{f}(x_1, x_2) = \sum_n \sum_m f(x_1 - k_1, x_2 - k_2)k(n, m), \tag{7.6}$$

where $n$ and $m$ are the indices of the spatial grid of the kernel function. Alternatively, one can view convolution as sliding the kernel over each pixel of the image and computing the dot product of overlapping kernel and pixel values as illustrated in Figure 7.3.

### Local connectivity and weight sharing

As mentioned earlier, the fully connected layer is only one of many possible connectivity patterns in a neural network. Two particular characteristics have been demonstrated to work particularly well with images:

- *Local connectivity*: each neuron is only connected to a small fraction of the input data.
- *Weight-sharing*: neurons share a common set of weights for the computation of their response.

This is partially motivated by the organisation of the animal visual cortex, where groups of neurons are only sensitive to a restricted part of the visual field. Translated into mathematical formalism, let's consider $n$ neurons and an input $\mathbf{x} \in \mathbb{R}^n$. Each neuron is connected only to a local neighbourhood of $K+1$ input elements around a particular input location, the $n$-th element of the input is connected to the $n$-th neuron, and all neurons share the same $K+1$ weights, then the output of that neuron can be computed by:

$$o(n) = \sum_{k=-K/2}^{K/2} w_k x_{i-k} \tag{7.7}$$

This corresponds to a 1-D convolution between a discrete signal and a kernel with kernel size $K+1$. For two-dimensional images, a 2D-convolution can be applied as described in 7.6. Convolutions with 2D kernels form the basis for *Convolutional Neural Networks* or CNNs LeCun et al. [1998], where many kernels are convolved with the input image, resulting in a bank of *filter responses* or *feature maps*.

### The convolutional layer

A convolutional layer consists of several filter kernels, each of which is convolved with all incoming features (3 in the very first layer for RGB colour-images). The results of the convolution for each incoming channel are summed element-wise to form the final output activation map

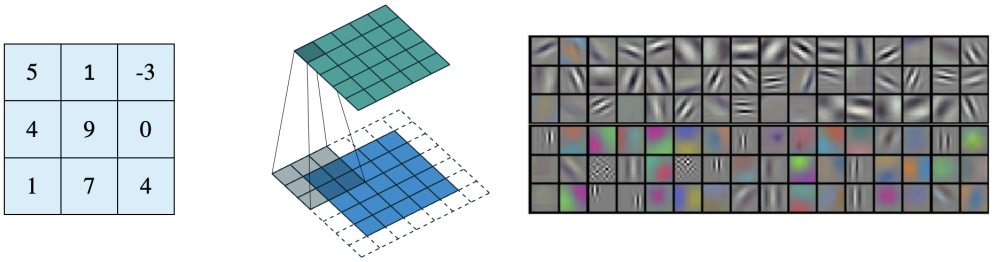



Figure 7.2: Left: example of a 3x3 kernel with associated weights. Middle: visualised 2d-convolution: the 3x3 kernel (grey) is slid over the input (blue) and each entry in the resulting feature map (turquoise) is the dot product of input values and kernel weights. Right: 96 11x11 filters learned by a convolutional neural network Krizhevsky et al. [2012]
.

for each filter kernel. To illustrate, a convolutional layer with 64 filter kernels will produce 64 activation maps.

## 7.2 Optimisation

Besides novel architectural designs including the use of ReLU Jarrett et al. [2009] activation functions and convolutional filters, strong and efficient optimisation techniques are another essential element for the recent success of the deep learning paradigm.

### 7.2.1 Gradient descent

In order for neural networks to learn, a parameter-dependent scoring function $f_\theta : X \times Y \to \mathbb{R}$ can be defined that maps data from an input space $X$ and an output from a model belonging to the output space $Y$ to a scalar value. The goal of optimisation is to find parameters $\theta$ that minimise that function. If $x$ and $y$ come from a probability distribution, this equates to minimising the expected value over the distribution of associated costs for all samples $x$ and $y$:

$$\theta^* = \arg\min_\theta \mathbb{E}_{x,y}[f_\theta(x, y, \theta)] \tag{7.8}$$

In most real-world applications a closed-form analytical solution is not available, and one needs

to resort to iterative, approximating approaches. One such approach is gradient descent, where one takes many steps $t$ in the opposite direction of the gradient (meaning updating the parameters by subtracting the gradient with respect to the scoring function), which points along the steepest descent of the loss curve, or generally speaking, surface:

$$\theta_{t+1} = \theta_t - \gamma \nabla_\theta f \tag{7.9}$$

$\gamma$ is referred to as the step size or learning rate and is a hyperparameter that determines the magnitude of the directional vector along which to move during optimisation.

### Stochastic gradient descent

In equation 7.9, the gradient is computed with respect to all samples in the data, which quickly becomes computationally infeasible for large datasets. Instead, one can compute the gradient with respect to a subsample, or batch, of the data. While this decreases the computational load, it also increases the chances of escaping local minima, which are common on the non-convex and high-dimensional loss surfaces of neural networks, due to the inherent noise that comes from estimating the next step from a small subsample.

### Momentum

Due to the non-convex nature of neural network optimisation, an additional momentum term can be introduced to increase chances of convergence towards a global minimum instead of becoming trapped in sub-optimal, local minima. This is achieved by simply adding the directional previous update to the current one. Intuitively, this is analogous to an object travelling down an inclined surface and picking up speed, or momentum, due to gravity. Once the lowest point in the neighbourhood (or a local minimum of the cost function in our picture) has been reached, the object will continue to travel upwards against gravity for a certain amount of time until stopping or reaching the next inclined terrain.

$$\Delta\theta_t = \alpha\Delta\theta_{t-1} - \gamma\nabla_\theta f \tag{7.10}$$

$$\theta_{t+1} = \theta_t + \Delta\theta_t, \tag{7.11}$$

where $\alpha \in [0, 1]$ is an another hyperparameter that controls the strength of the momentum term.

#### Adam

Adam Kingma and Ba [2014] is a very popular momentum-based optimiser for neural networks that keeps track of previous updates via an exponentially decaying moving-average:

$$g_t = \nabla_\theta f \tag{7.12}$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{7.13}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{7.14}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{7.15}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{7.16}$$

$$\theta_{t+1} = \theta_t - \gamma \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{7.17}$$

where $\beta_1$ and $\beta_2 \in [0, 1]$ control the degree of exponential decay, $\gamma$ is the learning rate and $\epsilon$ is a small value to avoid division by zero.

### 7.2.2 Backpropagation and automatic differentiation

The parameters $\theta$ to minimise the loss function are the weights defined in Equation 7.3, or, in case of a convolutional neural network, the elements of the kernel $\mathbf{k}$ as displayed in Figure 7.3. The fundamental learning process of neural networks is that these weights are adjusted via gradient descent, and the value by which weight changes per iteration depends on the partial

derivative of the final loss function with respect to that weight. We can recall from equation 7.4 that a neural network can be viewed as a concatenation of individual functions per layer. To calculate a specific gradient for one of these inner functions, one needs to apply the chain-rule for differentiation, which involves computation of derivatives of the outer layers as well. The repeated application of the chain rule to arrive at a certain location within the network is referred to as *backpropagation* Rumelhart et al. [1986] and is the key mathematical concept behind neural network optimisation. Modern deep learning software frameworks provide automatic backpropagation with the computation of partial derivatives in an efficient way also referred to as automatic differentiation.

### 7.2.3 Other common components of neural networks

#### Down-and upsampling

In computer vision applications, neural networks often map a high dimensional input such as an image to either a scalar as in classification or another image as in style transfer, image translation or image segmentation. A common element of any convolutional filter-based network architecture is a change of spatial resolution of the input accompanied by an increase in the amount of feature maps from applying the learned filters. This change can be achieved either via various image interpolation methods, pooling or strided convolutions, where the filter response is only calculated for every $n$-th pixel in the image. When applying a filter with a stride of 2 and appropriate padding for instance, the resulting feature map has only half the resolution.

#### Max- and average-pooling

Another commonly applied strategy for sub-sampling images is pooling. In *max-pooling* Yamaguchi et al. [1990], a two-dimensional filter window (usually a square grid of varying sizes) is slid over the input similarly to convolution, while only the maximum value from the input values within that window is propagated, resulting in reduced dimensionality depending on

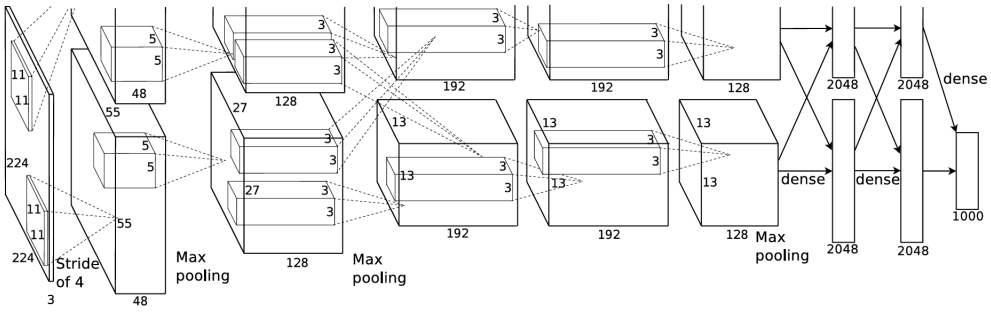



Figure 7.3: Schematic overview of the convolutional neural network architecture from Krizhevsky et al. [2012]. Height and width of the blocks correspond to the spatial resolution of input and features across the network, while depth shows the number of feature or activation maps. Various pooling operations throughout the network, reshaping and fully-connected layers towards the end transform the image to a low-dimensional vector denoting scores for a number of predicted classes.

the window size. *average pooling* works analogously with the arithmetic mean instead of the maximum.

### 7.2.4 Residual networks

Works including Simonyan and Zisserman [2014] and Szegedy et al. [2014] have demonstrated that increased neural network depth, *i.e.* number of layers, leads to better performance in image classification tasks. However, adding more and more layers eventually leads to saturation and an actual decrease in performance He et al. [2015]. Empirically, adding so called short-cut, skip or residual connections within the network helps to stabilise training and convergence even for very deep convolutional architectures with 50 layers and more. Li et al. [2017] make use of various techniques to visualise the topology of loss surfaces of different architectures and note striking differences between deep networks with and without skip-connections in terms of smoothness (absence of many large peaks and valleys), which suggests better conditions for optimisation via gradient descent.

Mathematically, skip connections simply refer to adding the original input $\mathbf{x}$ back to the output of an arbitrary amount of layers, denoted by a function $f$:

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{x} \tag{7.18}$$

Note that the addition is performed element-wise. In case of a dimensional mismatch between $\mathbf{x}$ and $f(\mathbf{x})$, $\mathbf{x}$ can be mapped to the same dimensions via a shallower function (*i.e.* less parameters or overall layers than $f$), often implemented via a convolution with a kernel size of 1. Skip-connections are present in a wide range of modern deep learning architectures.

# Bibliography


Labelbox. *Online, Available: https://labelbox.com*, 2021.

H. Abbasi, S. Mollet, S. A. Williams, L. J. Lim, M. R. Battin, T. F. Besier, and A. J. C. McMorland. Deep-learning for automated markerless tracking of infants general movements. *bioRxiv*, 2022. URL https://api.semanticscholar.org/CorpusID:250645976.

R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4431–4440, 2019. URL https://api.semanticscholar.org/CorpusID:102350964.

J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440, 1999. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1998.0744. URL https://www.sciencedirect.com/science/article/pii/S1077314298907445.

M. Airaksinen, O. Räsänen, E. Ilén, T. Häyrinen, A. Kivi, V. Marchi, A. Gallen, S. Blom, A. Varhe, N. Kaartinen, L. Haataja, and S. Vanhatalo. Automatic posture and movement tracking of infants with wearable movement sensors. *Scientific Reports*, 2020.

Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18490–18500, 2021. URL https://api.semanticscholar.org/CorpusID:244729249.

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1021, 2009a. URL https://api.semanticscholar.org/CorpusID:1430002.

M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE, 2009b.

M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. *ACM SIGGRAPH 2005 Papers*, 2005. URL `https://api.semanticscholar.org/CorpusID:3423879`.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *ArXiv*, abs/1701.07875, 2017. URL `https://api.semanticscholar.org/CorpusID:13943041`.

T. Bae, K. Choi, and M. Mun. Level walking and stair climbing gait in above-knee amputees. *Journal of medical engineering & technology*, 33(2):130–135, 2009.

V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 468–475. IEEE, 2017.

A. Bielski and P. Favaro. Emergence of object segmentation in perturbed generative models. *ArXiv*, abs/1905.12663, 2019.

J. Bjorck, C. P. Gomes, and B. Selman. Understanding batch normalization. In *Neural Information Processing Systems*, 2018. URL `https://api.semanticscholar.org/CorpusID:46956723`.

A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

M. Caron, H. Touvron, I. Misra, H. J'egou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. URL https://api.semanticscholar.org/CorpusID:233444273.

J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5932–5941, 2018.

Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2018. URL https://api.semanticscholar.org/CorpusID:54457478.

T. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Comput. Vis. Image Underst.*, 61:38–59, 1995. URL https://api.semanticscholar.org/CorpusID:15242659.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:52967399.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *ArXiv*, abs/2105.05233, 2021.

H. Doi, A. Furui, R. Ueda, K. Shimatani, M. Yamamoto, K. Sakurai, C. Mori, and T. Tsuji. Spatiotemporal patterns of spontaneous movement in neonates are significantly linked to risk of autism spectrum disorders at 18 months old. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:261121995.

M. Dombrowski, H. Reynaud, M. Baugh, and B. Kainz. Foreground-background separation

through concept distillation from generative image foundation models. 2022. URL https://api.semanticscholar.org/CorpusID:255340489.

C. Einspieler and H. F. R. Prechtl. Prechtl's assessment of general movements: a diagnostic tool for the functional assessment of the young nervous system. *Mental retardation and developmental disabilities research reviews*, 11 1:61–7, 2005. URL https://api.semanticscholar.org/CorpusID:13519571.

C. Einspieler, P. B. Marschik, A. F. Bos, F. Ferrari, G. Cioni, and H. F. R. Prechtl. Early markers for cerebral palsy: insights from the assessment of general movements. *Future Neurology*, 7:709–717, 2012. URL https://api.semanticscholar.org/CorpusID:72978617.

C. Einspieler, P. B. Marschik, J. Pansy, A. Scheuchenegger, M. Krieber, H. Yang, M. K. Kornacka, E. Rowinska, M. Soloveichick, and A. F. Bos. The general movement optimality score: a detailed assessment of general movements during preterm and term age. *Developmental Medicine & Child Neurology*, 58, 2016. URL https://api.semanticscholar.org/CorpusID:205068790.

D. Epstein, T. Park, R. Zhang, E. Shechtman, and A. A. Efros. Blobgan: Spatially disentangled scene representations. *ArXiv*, abs/2205.02837, 2022.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2004. URL https://api.semanticscholar.org/CorpusID:2277383.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

F. Ferrari, G. Cioni, C. Einspieler, M. Roversi, A. F. Bos, P. B. Paolicelli, A. Ranzi, and H. F. R. Prechtl. Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy. *Archives of pediatrics & adolescent medicine*, 156 5:460–7, 2002a. URL https://api.semanticscholar.org/CorpusID:16257635.

F. Ferrari, G. Cioni, C. Einspieler, M. F. Roversi, A. F. Bos, P. B. Paolicelli, A. Ranzi, and

H. R. F. Prechtl. Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy. *Archives of pediatrics & adolescent medicine*, 156 5:460–7, 2002b.

C. Ferraris, V. Cimolin, L. Vismara, V. Votta, G. Amprimo, R. Cremascoli, M. Galli, R. Nerino, A. Mauro, and L. Priano. Monitoring of gait parameters in post-stroke individuals: A feasibility study using rgb-d sensors. *Sensors*, 21(17), 2021. ISSN 1424-8220. doi: 10.3390/s21175945. URL https://www.mdpi.com/1424-8220/21/17/5945.

M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, C-22:67–92, 1973. URL https://api.semanticscholar.org/CorpusID:14554383.

O. Freifeld, A. Weiss, S. Zuffi, and M. J. Black. Contour people: A parameterized model of 2d articulated human shape. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 639–646, 2010. URL https://api.semanticscholar.org/CorpusID:936747.

L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. URL https://api.semanticscholar.org/CorpusID:206593710.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014a. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

I. J. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *ArXiv*, abs/1701.00160, 2016. URL https://api.semanticscholar.org/CorpusID:5216145.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014b.

R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in

the wild. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. URL https://api.semanticscholar.org/CorpusID:13637778.

M. Hadders-Algra. Early diagnosis and early intervention in cerebral palsy. *Frontiers in Neurology*, 5, 2014. URL https://api.semanticscholar.org/CorpusID:14614944.

L. Han, S. Musunuri, M. R. Min, R. Gao, Y. Tian, and D. N. Metaxas. Ae-stylegan: Improved training of style-based auto-encoders. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 955–964, 2021. URL https://api.semanticscholar.org/CorpusID:239016070.

S. Han and S. Lee. A vision-based motion capture and recognition framework for behavior-based safety management. *Automation in Construction*, 35:131–141, 2013. URL https://api.semanticscholar.org/CorpusID:111322276.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016a.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016b.

X. He, B. Wandt, and H. Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image generation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1215–1225, 2021.

X. He, B. Wandt, and H. Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. *ArXiv*, abs/2205.10636, 2022.

N. Hesse, G. Stachowiak, T. Breuer, and M. Arens. Estimating body pose of infants in depth

images using random ferns. *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

N. Hesse, C. Bodensteiner, M. Arens, U. G. Hofmann, R. Weinberger, and A. S. Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *European Conference on Computer Vision*, pages 32–49. Springer, 2018a.

N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Müller-Felber, and A. S. Schroeder. Learning an infant body model from RGB-D data for accurate full body motion analysis. 2018b.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

M. R. I. Hossain and J. J. Little. Exploiting temporal information for 3d human pose estimation. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 69–86. Springer, 2018.

X. Huang and S. J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. URL `https://api.semanticscholar.org/CorpusID:6576859`.

E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014a.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014b.

U. Iqbal, A. Doering, H. Yasin, B. Krüger, A. Weber, and J. Gall. A dual-source approach for 3d human pose estimation from a single image. *Comput. Vis. Image Underst.*, 172:37–49, 2018.

A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *ECCV Workshops*, 2018. URL `https://api.semanticscholar.org/CorpusID:52191460`.

T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4016–4027. Curran Associates, Inc., 2018. URL `http://papers.nips.cc/paper/7657-unsupervised-learning-of-object-landmarks-through-conditional-image-generation.pdf`.

T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Learning landmarks from unaligned data using image translation. *ArXiv*, abs/1907.02055, 2019.

T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8784–8794, 2020a. URL `https://api.semanticscholar.org/CorpusID:215764116`.

T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8784–8794, 2020b.

K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009. doi: 10.1109/ICCV.2009.5459469.

B. Jiang, Y. Hong, H. Bao, and J. Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. *ArXiv*, abs/2201.12792, 2022.

J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *ArXiv*, abs/1603.08155, 2016. URL `https://api.semanticscholar.org/CorpusID:980236`.

A. Kanazawa, D. W. Jacobs, and M. Chandraker. Warpnet: Weakly supervised matching for single-view reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3253–3261, 2016.

A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131. IEEE Computer Society, 2018.

D. Karch, K.-S. Kim, K. Wochner, J. Pietz, H. Dickhaus, and H. Philippi. Quantification of the segmental kinematics of spontaneous infant movements. *Journal of biomechanics*, 41(13), September 2008. doi: 10.1016/j.jbiomech.2008.06.033.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. URL `https://api.semanticscholar.org/CorpusID:54482423`.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019a.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2019b. URL `https://api.semanticscholar.org/CorpusID:209202273`.

I. Katircioglu, B. Tekin, M. Salzmann, V. Lepetit, and P. Fua. Learning latent representations of 3d human pose with deep neural networks. *Int. J. Comput. Vis.*, 126(12):1326–1341, 2018.

M. H. Khan, M. Schneider, M. S. Farid, and M. Grzegorzek. Detection of infantile movement disorders in video data using deformable part-based model. *Sensors*, 18, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

M. Kocabas, N. Athanasiou, and M. J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

J. M. Kohler, H. Daneshmand, A. Lucchi, T. Hofmann, M. Zhou, and K. Neymeyr. Exponential convergence rates for batch normalization: The power of length-direction decoupling in non-convex optimization. In *International Conference on Artificial Intelligence and Statistics*, 2018. URL https://api.semanticscholar.org/CorpusID:52941550.

N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

J. Kondragunta and G. Hirtz. Gait parameter estimation of elderly people using 3d human pose estimation in early detection of dementia. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5798–5801, 2020. URL https://api.semanticscholar.org/CorpusID:221388327.

P. Kontschieder, J. F. Dorn, C. Morrison, R. Corish, D. Zikic, A. Sellen, M. D'Souza, C. P. Kamm, J. Burggraaff, P. Tewarie, et al. Quantifying progression of multiple sclerosis via classification of depth videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 429–437. Springer, 2014.

J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane. *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*, pages 151–170. Springer Netherlands, Dordrecht, 1996. ISBN 978-94-009-0279-4. doi: 10.1007/978-94-009-0279-4_9. URL https://doi.org/10.1007/978-94-009-0279-4_9.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. URL https://api.semanticscholar.org/CorpusID:195908774.

J. N. Kundu, S. Seth, V. Jampani, M. Rakesh, R. Babu, and A. Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6151–6161, 2020a.

J. N. Kundu, S. Seth, V. Jampani, M. Rakesh, R. V. Babu, and A. Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6151–6161, 2020b.

H. Kwak and B.-T. Zhang. Generating images part by part with composite generative adversarial networks. *ArXiv*, abs/1607.05387, 2016.

Y. Kwon, D. Kim, D. Ceylan, and H. Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *ArXiv*, abs/2109.07448, 2021.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:2278–2324, 1998. URL https://api.semanticscholar.org/CorpusID:14542261.

H. Li, Z. Xu, G. Taylor, and T. Goldstein. Visualizing the loss landscape of neural nets. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:3693334.

Y. Li, K. Li, S. Jiang, Z. Zhang, C. Huang, and R. Y. D. Xu. Geometry-driven self-supervised method for 3d human pose estimation. volume 34, pages 11442–11449, Apr. 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/6808.

M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3d pose sequence machines. *CoRR*, abs/1707.09695, 2017. URL http://arxiv.org/abs/1707.09695.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, e. D. Zitnick, C. Lawrence", T. Pajdla, B. Schiele, and T. Tuytelaars. Microsoft coco: Common objects in

context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016. URL https://api.semanticscholar.org/CorpusID:10716717.

S. Liu, W. Chen, T. Li, and H. Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *ArXiv*, abs/1901.05567, 2019. URL https://api.semanticscholar.org/CorpusID:58014105.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2015a. URL https://api.semanticscholar.org/CorpusID:5328073.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015b.

M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015c.

D. Lorenz, L. Bereska, T. Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10947–10956, 2019a. URL https://api.semanticscholar.org/CorpusID:81979136.

D. Lorenz, L. Bereska, T. Milbich, and B. Ommer. Unsupervised part-based disentangling of object shape and appearance. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10947–10956, 2019b.

S. E. A. L. T. A. Luca Schmidtke, Athanasios Vlontzos and B. Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2016. URL https://api.semanticscholar.org/CorpusID:206771128.

J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings IEEE International Conference on Computer Vision (ICCV)*, Piscataway, NJ, USA, Oct. 2017a. IEEE.

J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017b.

K. D. McCay, E. S. L. Ho, H. P. H. Shum, G. Fehringer, C. Marcroft, and N. D. Embleton. Abnormal infant movements classification with deep learning on pose-based features. *IEEE Access*, 8:51582–51592, 2020. doi: 10.1109/ACCESS.2020.2980269.

R. Mehrizi, X. Peng, Z. Tang, X. Xu, D. N. Metaxas, and K. Li. Toward marker-free 3d pose estimation in lifting: A deep multi-view solution. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 485–491, 2018. URL https://api.semanticscholar.org/CorpusID:3634386.

D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

G. E. Mena, D. Belanger, S. W. Linderman, and J. Snoek. Learning latent permutations with gumbel-sinkhorn networks. *ArXiv*, abs/1802.08665, 2018. URL https://api.semanticscholar.org/CorpusID:3515469.

L. M. Mescheder, A. Geiger, and S. Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, 2018a. URL https://api.semanticscholar.org/CorpusID:3345317.

L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on*

*Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2018b. URL https://api.semanticscholar.org/CorpusID:54465161.

L. Metz, B. Poole, D. Pfau, and J. N. Sohl-Dickstein. Unrolled generative adversarial networks. *ArXiv*, abs/1611.02163, 2016. URL https://api.semanticscholar.org/CorpusID:6610705.

B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65:99–106, 2020. URL https://api.semanticscholar.org/CorpusID:213175590.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. *ArXiv*, abs/1802.05957, 2018. URL https://api.semanticscholar.org/CorpusID:3366315.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016a. URL https://api.semanticscholar.org/CorpusID:13613792.

A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016b. Springer International Publishing. ISBN 978-3-319-46484-8.

T. D. Nguyen, T. Le, H. T. Vu, and D. Q. Phung. Dual discriminator generative adversarial nets. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:1059659.

I. Novak, C. Morgan, L. Adde, J. A. Blackman, R. N. Boyd, J. E. Brunstrom-Hernandez, G. Cioni, D. L. Damiano, J. Darrah, A.-C. Eliasson, L. S. de Vries, C. Einspieler, M. C. Fahey, D. L. Fehlings, D. M. Ferriero, L. K. Fetters, S. Fiori, H. Forssberg, A. M. Gordon, S. Greaves, A. Guzzetta, M. Hadders-Algra, R. T. Harbourne, A. Kakooza-Mwesige, P. Karlsson, L. Krumlinde-Sundholm, B. Latal, A. Loughran-Fowlds, N. L. Maitre, S. Mcintyre, G. Noritz, L. Pennington, D. M. Romeo, R. B. Shepherd, A. J. Spittle, M. Thornton,

J. Valentine, K. Walker, R. White, and N. Badawi. Early, accurate diagnosis and early intervention in cerebral palsy: Advances in diagnosis and treatment. *JAMA Pediatrics*, 171: 897–907, 2017. URL https://api.semanticscholar.org/CorpusID:205141458.

M. D. Olsen, A. Herskind, J. B. Nielsen, and R. R. Paulsen. Model-based motion tracking of infants. 2014.

W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2329–2336, 2014.

J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019a. URL https://api.semanticscholar.org/CorpusID:58007025.

T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2332–2341, 2019b.

T. Parr and K. J. Friston. The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience*, 12, 2018. URL https://api.semanticscholar.org/CorpusID:53281608.

G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017.

D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

H. F. R. Prechtl. Qualitative changes of spontaneous movements in fetus and preterm infant are a marker of neurological dysfunction. *Early human development*, 23 3:151–8, 1990. URL `https://api.semanticscholar.org/CorpusID:29304873`.

H. F. R. Prechtl. General movement assessment as a method of developmental neurology: new paradigms and their consequences. the 1999 ronnie mackeith lecture. *Developmental Medicine & Child Neurology*, 43:836–842, 2001. URL `https://api.semanticscholar.org/CorpusID:146695230`.

H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng. Cross view fusion for 3d human pose estimation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4341–4350, 2019.

A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. URL `https://api.semanticscholar.org/CorpusID:11758569`.

H. Rahmati, R. Dragon, O. M. Aamo, L. Adde, Ø. Stavdahl, and L. V. Gool. Weakly supervised motion segmentation with particle matching. *Computer Vision and Image Understanding*, 2015.

V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *European Conference on Computer Vision*, pages 33–47. Springer, 2014.

N. Ravi, J. Reizenstein, D. Novotny, T. Gordon, W.-Y. Lo, J. Johnson, and G. Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

S. H. Rezatofighi, R. Kaskman, F. T. Motlagh, J. Q. Shi, D. Cremers, L. Leal-Taixé, and I. D. Reid. Deep perm-set net: Learn to predict sets with unknown permuta-

tion and cardinality using deep neural networks. *ArXiv*, abs/1805.00613, 2018. URL https://api.semanticscholar.org/CorpusID:13744272.

H. Rhodin, M. Salzmann, and P. V. Fua. Unsupervised geometry-aware representation for 3d human pose estimation. *ArXiv*, abs/1804.01110, 2018a.

H. Rhodin, J. Spörri, I. Katircioglu, V. Constantin, F. Meyer, E. Müller, M. Salzmann, and P. V. Fua. Learning monocular 3d human pose estimation from multi-view images. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018b.

E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2287–2296, 2020. URL https://api.semanticscholar.org/CorpusID:220936362.

I. Rocco, R. Arandjelovic, and J. Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. URL https://api.semanticscholar.org/CorpusID:245335280.

O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. URL https://api.semanticscholar.org/CorpusID:3719281.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. URL https://api.semanticscholar.org/CorpusID:205001834.

S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry. How does batch normalization help optimization? (no, it is not about internal covariate shift). In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:53104146.

B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 422–429. IEEE, 2010.

L. Schmidtke, A. Vlontzos, S. Ellershaw, A. Lukens, T. Arichi, and B. Kainz. Unsupervised human pose estimation through transforming shape templates. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2484–2494, 2021.

L. Schmidtke, B. Hou, A. Vlontzos, and B. Kainz. Self-supervised 3d human pose estimation in static video via neural rendering. *ArXiv*, abs/2210.04514, 2022. URL https://api.semanticscholar.org/CorpusID:252780712.

J. Shotton, T. Sharp, A. Kipman, A. W. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *CVPR 2011*, pages 1297–1304, 2011. URL https://api.semanticscholar.org/CorpusID:7731948.

Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL https://api.semanticscholar.org/CorpusID:14124313.

K. K. Singh, U. Ojha, and Y. J. Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6483–6492, 2019.

A. Stahl, C. Schellewald, O. Stavdahl, O. M. Aamo, L. Adde, and H. Kirkerød. An optical flow-based method to predict infantile cerebral palsy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20:605–614, 2012.

J. Stenum, K. M. Cherry-Allen, C. O. Pyles, R. Reetzke, M. F. Vignos, and R. T. Roemmich. Applications of pose estimation in human health and performance across the lifespan. *Sensors (Basel, Switzerland)*, 21, 2021. URL https://api.semanticscholar.org/CorpusID:243794363.

K. N. Stevens. Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 32:47–55, 1960. URL https://api.semanticscholar.org/CorpusID:119755920.

S.-Y. Su, F. Yu, M. Zollhoefer, and H. Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. 2021.

M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2014. URL https://api.semanticscholar.org/CorpusID:206592484.

P. Teitelbaum, O. Teitelbaum, J. Nye, J. Fryman, and R. G. Maurer. Movement analysis in infancy may be useful for early diagnosis of autism. *Proceedings of the National Academy of Sciences of the United States of America*, 95 23:13982–7, 1998a. URL https://api.semanticscholar.org/CorpusID:21412445.

P. Teitelbaum, O. Teitelbaum, J. Nye, J. B. Fryman, and R. G. Maurer. Movement analysis in infancy may be useful for early diagnosis of autism. *Proceedings of the National Academy of Sciences of the United States of America*, 95 23:13982–7, 1998b.

J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3229–3238, 2017a. URL https://api.semanticscholar.org/CorpusID:19833076.

J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. In *Advances in neural information processing systems*, pages 844–855, 2017b.

J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6361–6371, 2019.

D. Tomè, M. Toso, L. de Agapito, and C. Russell. Rethinking pose in 3d: Multi-stage refinement and recovery for markerless motion capture. *2018 International Conference on 3D Vision (3DV)*, pages 474–483, 2018.

J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1799–1807. Curran Associates, Inc., 2014. URL `http://papers.nips.cc/paper/5573-joint-training-of-a-convolutional-network-and-a-graphical-model-for-human-pose-e` `pdf`.

A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

P. Upchurch, J. V. Gardner, G. Pleiss, K. Bala, R. Pless, N. Snavely, and K. Q. Weinberger. Deep feature interpolation for image content changes. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6090–6099, 2016. URL `https://api.semanticscholar.org/CorpusID:19009954`.

A. van den Oord, N. Kalchbrenner, L. Espeholt, K. Kavukcuoglu, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. *ArXiv*, abs/1606.05328, 2016. URL `https://api.semanticscholar.org/CorpusID:14989939`.

G. Varol, D. Ceylan, B. C. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3d human body shapes. *ArXiv*, abs/1804.04875, 2018. URL `https://api.semanticscholar.org/CorpusID:4898805`.

S. Wang, K. Schwarz, A. Geiger, and S. Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. URL `https://api.semanticscholar.org/CorpusID:252968216`.

S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. *ArXiv*, abs/1804.06208, 2018a. URL `https://api.semanticscholar.org/CorpusID:4934594`.

B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018b.

J. Xu, Z. Yu, B. Ni, J. Yang, X. Yang, and W. Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

K. Yamaguchi, K. Sakamoto, T. Akabane, and Y. Fujimoto. A neural network for speaker-independent isolated word recognition. In *ICSLP*, 1990. URL `https://api.semanticscholar.org/CorpusID:41281636`.

J. Yang, A. Kannan, D. Batra, and D. Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *ArXiv*, abs/1703.01560, 2017.

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR 2011*, pages 1385–1392. IEEE, 2011.

Y. Yang, H. Bilen, Q. Zou, W. Y. Cheung, and X.-W. Ji. Learning foreground-background segmentation from improved layered gans. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 366–375, 2022.

Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018a. URL `https://api.semanticscholar.org/CorpusID:4810698`.

Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee. Unsupervised discovery of object landmarks as structural representations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018b.

Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019. URL `https://api.semanticscholar.org/CorpusID:80628414`.

Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2018. URL `https://api.semanticscholar.org/CorpusID:56178817`.

S. Zuffi and M. J. Black. The stitched puppet: A graphical model of 3D human shape and pose. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 3537–3546, June 2015.

S. Zuffi, O. Freifeld, and M. J. Black. From pictorial structures to deformable structures. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3546–3553. IEEE, June 2012.