DEGREE PROJECT IN TECHNOLOGY,
SECOND CYCLE, 30 CREDITS
STOCKHOLM, SWEDEN 2020

3D Human Pose Estimation

KTH Thesis Report
Data, Methodology and Results
Draft

Sri Datta Budaraju

KTH ROYAL INSTITUTE OF TECHNOLOGY
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Authors

Sri Datta Budaraju <budaraju@kth.se>
School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology

Place for Project

Stockholm, Sweden
Stuttgart, Germany

Examiner

Danica Kragic Jensfelt
Stockholm, Sweden
KTH Royal Institute of Technology

Supervisor

Hedvig Kjellström
Stockholm, Sweden
KTH Royal Institute of Technology

Supervisor - Host

Arij Bouaziz

Stuttgart, Germany

Mercedes-Benz AG, Research and Development

Abstract

This is a template for writing thesis reports for the ICT school at KTH. I do not own any of the images provided in the template and this can only be used to submit thesis work for KTH.

The report needs to be compiled using XeLaTeX as different fonts are needed for the project to look like the original report. You might have to change this manually in overleaf.

This template was created by Hannes Rabo <hannes.rabo@gmail.com or hrabo@kth.se> from the template provided by KTH. You can send me an email if you need help in making it work for you.

Write an abstract. Introduce the subject area for the project and describe the problems that are solved and described in the thesis. Present how the problems have been solved, methods used and present results for the project. Use probably one sentence for each chapter in the final report.

The presentation of the results should be the main part of the abstract. Use about ½ A4-page. English abstract

Keywords

Template, Thesis, Keywords ...

Abstract

TO BE WRITTEN
Svenskt abstract Svensk version av abstract – samma titel på svenska som på engelska.

Skriv samma abstract på svenska. Introducera ämnet för projektet och beskriv problemen som löses i materialet. Presentera

Nyckelord

Kandidat examensarbete, ...

Acknowledgements

Write a short acknowledgements. Don't forget to give some credit to the examiner and supervisor.

Acronyms

ANN Artificial Neural Network

AR/VR Augmented Reality/Virtual Reality

HPE Human Pose Estimation

KLD Kullback–Leibler Divergence

L1 Least Absolute Deviations

MMVAE Mixture-of-Experts Multimodal Variational Auto-Enocder

MPJPE Mean Per Joint Position Estimate

MSE Mean Squared Error

MVAE Multimodal Variational Auto-Enocder

MoCap Motion Capture

NRSfM Non-Rigid Structure from Motion

POV Point of View

RGB Red Green Blue

SOTA State-of-The-Art

UMAP Uniform Manifold Approximation and Projection

VAE Variational Auto-Encoder

GAN Generative Adversarial Network

WGAN Wasserstein Generative Adversarial Network

WGAN-GP Wasserstein Generative Adversarial Network (WGAN) with Gradient Penalty

JS-divergence Jensen–Shannon Divergence

EM distance Earth Mover's Distance

β-VAE Beta Variational Auto-Encoder

Contents

Chapter 1

Introduction

With rapid advancements in deep learning facilitated by the developments in computational hardware, there has been tremendous growth in computer vision research and its applications [12]. One of the major tasks of computer vision that is required for real-world applications is to perceive and understand dynamic objects and more importantly humans.

Human Pose Estimation, also referred to as HPE is a fundamental problem in computer vision that also forms a basis for human action and gesture recognition as well as human motion prediction. Human Pose Estimation (HPE) is defined as the localization of human joints (also known as keypoints, including head, eyes, ears, nose etc) mainly in images and videos in either a 2D or 3D coordinate space. The widely available and used data like images and videos are 2-dimensional data and lack spatial information which is crucial for most of the applications like autonomous driving, Augmented Reality/Virtual Reality (AR/VR), social robots etc. Hence this thesis focus is on 3D Human Pose Estimation.

1.1 Background

There has been a lot of research done in 3D human pose estimation and more advancements have been made in the past few years leveraging the power of deep learning. The current state of the art methods explores various ways to solve the task using Red Green Blue (RGB)/Depth image channels, 2D poses, 3D poses, multi-view and sequential images. The main State-of-The-Art (SOTA) approaches either

directly estimate 3D pose from an RGB/D image in a bottom-up manner, or start with an intermediate 2D pose or, 2D joint heatmaps to finally recover the 3D pose by a *lifting* network. Typically, these approaches directly estimate the 3D coordinates of the keypoints or estimate the shape and camera parameters to reconstruct the 3D pose.

The former approaches are usually trained in a cascaded manner i.e by having an intermediate state that learns 2D pose in some way. Most of these methods have complex architectures that are hard to train or use multi-view images making it impractical to scale the training to the wild, where such data is very hard to obtain. Since 2D poses are naturally obtained by projecting 3D poses to a plane, the latter approach of lifting 2D-to-3D is an *ill-posed inverse* problem due to its inherent ambiguity.

Non-supervised (Weak/Semi/Self/Un-supervised) learning regimes have also gained traction in 3D HPE recently and many of the deep learning techniques that have already improved the results in other computer vision tasks (and even in supervised HPE), are yet to be explored.

1.2 Problem

How can we learn a strong visual representation of the data to tackle the 3D-pose estimation? Could data as its own supervisory goal (self-supervision) resolve the ambiguities of the pose estimation?

1.3 Goal

The main aim of the thesis is to investigate SOTA unsupervised learning approaches to estimate 3D human pose from images. And to also investigate 2D-to-3D lifting methods to tackle the challenges of 3D pose estimation in the wild.

Improvements in the aspects of ease of training procedure i.e requiring less data or less labor-intense labeling, inference speed, and most importantly accuracy is important and will directly impact its super tasks such as, action and gesture recognition, motion prediction and intention/behaviour prediction.

1.4 Benefits, Ethics and Sustainability

Human Pose Estimation plays a very important role to enable autonomous vehicles and robots to safely interact with humans. It plays a key role in developing higher dimensional communication platforms with AR/VR. It is crucial for surveillance systems to ensure public safety. However such important technologies are only as good as the intentions of its users. Mass surveillance of citizens by their governments is a matter of debate.

1.5 Methodology

The problem of 3D HPE has 3 aspects to be addressed and explored.

The neural network: The architecture and the kind of neural network to be used. 3D poses can be predicted using regular linear neural network, or using various forms of autoencoder architectures. These models can use linear, convolutional or graph networks to learn the features. This thesis focus on exploring all the above-mentioned kinds of networks to solve the 3D HPE within the context of probabilistic inference models, typically Variational Auto-Encoder (VAE)s, as a deterministic approach for an inherently ill-posed problem is not ideal.

The learning task: The model could either learn to directly predict the 3D coordinates of the keypoints, or learn structural parameters that could model a 3D pose. The thesis only explores the former task.

The learning technique (or the cost): . The model can be either trained by directly comparing the predicted 3D pose and the ground truth thus requiring 3D annotations or, by projecting the prediction back to 2D to compare with the input (requires only 2D annotations that could be acquire from SOTA in 2D HPE). Adversarial training and self-supervision techniques have also given promising results in the last couple of years. The thesis's prime focus is on the first aspect of investigating the merit in using VAE for 3D HPE hence direct comparison of 3D pose and ground truth would be used and could be further extended to other techniques with moderate modifications.

1.6 Stakeholders

Daimler's 'Environment Perception for Autonomous Driving' R&D team in Stuttgart conducts cutting-edge research in the field of Computer vision and Deep Learning to improve the State-Of-The-Art and to make Autonomous Driving a reality. This thesis is part of the team's on-going research in the area of Human Pose Estimation which would help autonomous cars better perceive, understand and interact with humans. Daimler/Mercedes-Benz autonomous cars try to understand humans both, inside and outside the car and Human Pose Estimation is a critical element to accomplish the task.

The question is also of interest to the research area of Human State/Action Recognition in specific and also to areas of computer graphics to model humans in 3D space. Hence it is beneficial to various areas that try to understand and interact with humans. The scientific communities in the areas of Autonomous Driving, AR/VR, Motion Capture, Computer Graphics, and Human-Robot interaction could be interested in the contribution of this thesis.

1.7 Delimitations

This thesis focuses only on 3D pose estimation and not the intermediate 2D pose. Data collection is not part of the thesis study but uses only publicly available, widely used and benchmarked datasets.

1.8 Outline

The current version of the draft consists of two chapters alone. Chapter 2, the Theoretical Background further contains related works and would later include theoretical concepts that would be touched in the methodology.

Chapter 2

Theoretical Background

This chapter provides the theory essential to understand the major components of the thesis. Prior knowledge of Artificial Neural Networks [11] and fundamental concepts of Deep Learning [15] is assumed.

2.1 Preliminary Concepts

2.1.1 Autoencoder

Autoencoders are a variant of Artificial Neural Networks (ANNs), which are designed to learn an identity function that generates the input data sample back. The network has a bottleneck(z), dividing the network into two parts, an encoder and a decoder as illustrated in fig[2.1.1]. The first network learns to compress the high dimensional input data to a low dimensional intermediate representation, *latent representation*, at the bottleneck. While the second network learns to reconstruct the data from the latent distribution. Thus learning to efficiently compress the data.

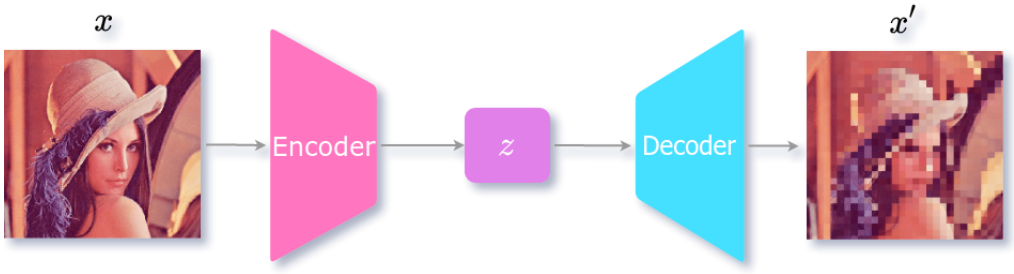

Figure 2.1.1: Illustration of autoencoder architecture.

To put it in other words, in the process of learning to reconstruct the data, the encoder learns to filter the most important features of the given data distribution, so as it preserves the complete properties within the limits of the bottleneck. While the decoder learns more general properties of the distribution which is used along with the compact latent representation from the encoder to fully recover the data distribution. The network is trained to minimize the similarity between the reconstruction and the original data sample. This similarity can be determined by metrics such as Mean Squared Error (MSE), Least Absolute Deviations (L1) or Cross-Entropy loss.

The idea of an autoencoder dates back to the '80s proposed as a method for pre-training and feature learning [5, 38], learned dimensionality reduction [22]. In recent years, autoencoders are most popularly used as generative networks leveraging their ability to learn feature representations in an unsupervised way. Another interesting variant of autoencoders is the denoising autoencoder [42], where the input is a noised data and the decoder generates original data without noise. This variant is further evolved to accomplish the tasks of image denoising, watermark removal, inpainting, super-resolution, colorization, de-colorization, and compression [48, 49]. The variant of autoencoders that is used in this thesis is introduced next.

2.1.2 Variational Autoencoders

VAEs, unlike standard autoencoders, learn to encode a data sample x as a probabilistic distribution rather than a deterministic value for the latent attribute [26]. The encoder q produces the probabilistic distribution by predicting two vectors that represent the mean μ and variance σ of distribution for each of the latent attributes of x. And the decoder p takes a random sample z from this distribution to recover the sample as illustrated in fig[2.1.2].

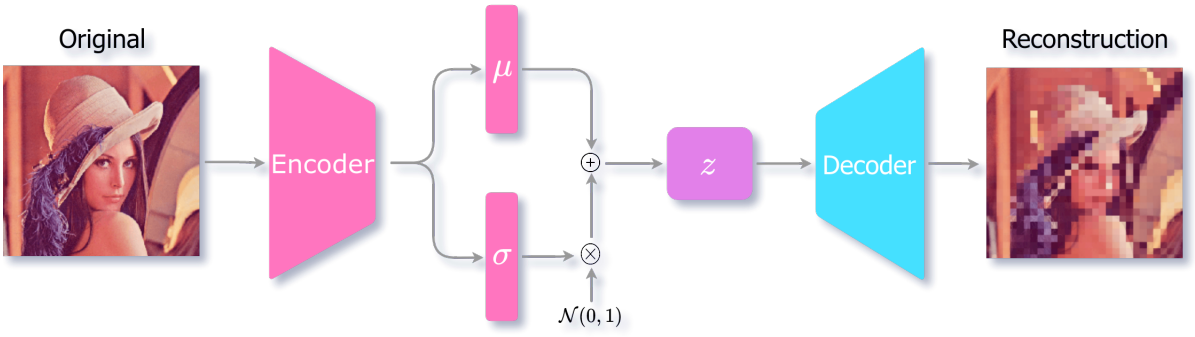

Figure 2.1.2: Illustration of Variational Autoencoder architecture.

To put it formally, we have a hidden variable z which generates x. Since we only have x and would want to learn z i.e $p(z|x)$. But computing this posterior distribution is hard as computing $p(x)$ Eq. 2.1 is usually itractable [27].

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$
$$p(x) = \int p(x|z)p(z)dz \tag{2.1}$$

Hence, we try to approximate the posterior distribution by another distribution $q(z|x)$ (the encoder) using variational inference. Variational inference uses optimzation to find a distribution that minimizes the Kullback–Leibler Divergence (KLD) to the posterior distribution, $\min D_{\mathrm{KL}}(q(z|x)\|p(z|x))$ while trying to keep the learnt distribution close to tthe true prior distribution $p(z)$[6]. The prior $p(z)$ is usually assumed to be a unit gaussian distribution. The above can also be achieved by maximizing:

$$\max \mathbb{E}_{q(z|x)} \log p(x|z) - D_{\mathrm{KL}}(q(z|x)||p(z)) \tag{2.2}$$

The first term is the above equation makes sure the reconstruction of the data sample x, while the second term tries to keep the learned distribution $q(z|x)$ close to the true prior $p(z)$. Hence the loss term to *minimize* while training the VAE is $\mathcal{L}_{\mathrm{VAE}}$ Eq. 2.3.

$$\mathcal{L}_{\mathrm{VAE}} = -\mathbb{E}_{q(z|x)} \log p(x|z) + D_{\mathrm{KL}}(q(z|x)||p(z)) = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{prior}} \tag{2.3}$$

However, the reconstruction error in the loss requires sampling z, which is a stochastic process and it is not possible to perform backpropagation. To address this problem, the ***reparametrization trick*** is used. Where ϵ is randomly sampled from a unit gaussian distribution $\mathcal{N}(0, 1)$ and is used to scale the variance σ of the latent distribution represented by the encoder $q_\theta(z|x)$. The sum of the mean μ and the scaled variance $\sigma \odot \epsilon$ gives z, which is now differentiable while being stochastic as illustrated in the fig[2.1.3].

Hence the learned latent space of a VAE is continuous while that of a standard autoencoder is discrete and clustered. As the decoder of the VAE is trained to generate data from this continuous space, it can generate realistic data by randomly sampling

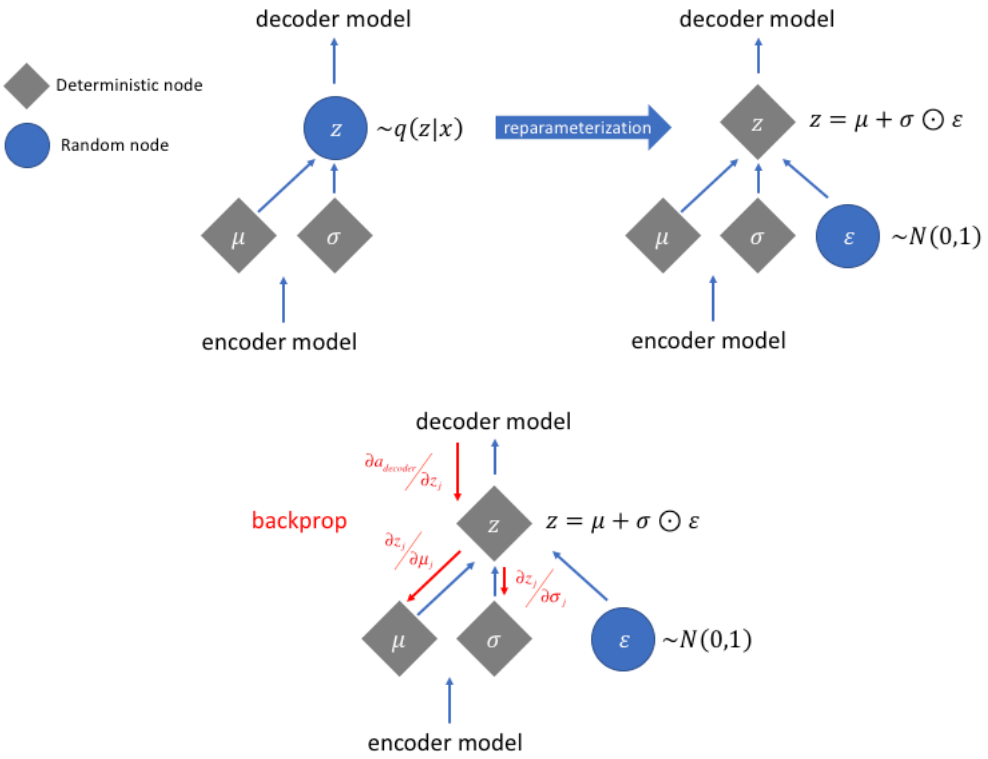

Figure 2.1.3: The reparametrization trick. Image Source [26]

from this infinitely large latent space as illustrated in the fig[2.1.4]. This also enables smooth interpolation of data produced from one point in the latent space to another. In addition to that, we can also perform arithmetics in vector space, similar to the popular example from Natural Language Processing, $King - Man + Women = Queen$ but on much higher dimensional embedding space.

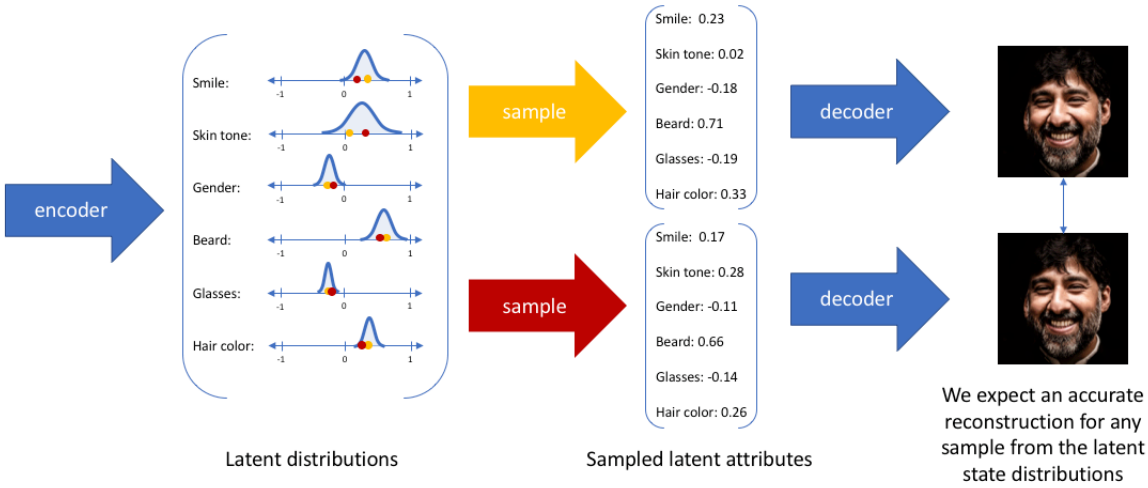

Figure 2.1.4: Probablistic distribution of latent attributes. Image Source [26]

2.1.3 Beta Variational Autoencoder

A VAE without the KLD term is effectively a standard autoencoder. As discussed the KLD term encourages the network to learn a distribution rather than a single value. If the variance of the distribution is not high, then it is again similar to an autoencoder. The more enforcement from KLD, the diverse the distribution. The VAE is forced to disentangle the representations, i.e the lesser is the correlation between each dimension in the latent space. Such disentangled representations are very useful for generative models. More importantly, it improves the interpretability of the latent space and can be leveraged to generalize to different downstream tasks. This emphasis on the latent space distributions can be achieved by disentangled variational autoencoders or Beta Variational Auto-Encoder (β-VAE), where β is the weight coefficient of the KLD term in the VAE loss function Eq.2.3. So, the loss for the β-VAE is $\mathcal{L}_{\text{VAE}}$ Eq. 2.4. The higher the beta the stronger the constrain on the disentanglement. This constraint will negatively affect the representation capability of the VAE.

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(z|x)} \log p(x|z) + \beta(D_{\text{KL}}(q(z|x)||p(z))) \tag{2.4}$$

2.1.4 Generative Adversarial Networks

Generative Adversarial Network (GAN) is an ANN that is used for generative tasks to make the prediction *realistic*. A GAN is a combination of 2 networks namely, the generator G and the discriminator D. The generator learns to map a random sample or say, noise Z drawn from a latent distribution with density p_z to a higher dimensional data distribution with density p_g. Where the discriminator takes the output of the generator and tries to differentiate real data samples x from the fakes which do not belong to the real distribution p_r as illustrated in the fig[2.1.5].

The goal of the generator is to produce samples, $G(z)$ that fool the discriminator into believing them as real samples. While the goal of the discriminator is to distinguish between the samples produced by the generator and the real samples by predicting reals as 1 and 0 for fakes. This inverse goals of the two networks can be viewed as a game of tug of war or a 2 player minimax game. The result of the game is that the generator would ideally learn to produce realistic data samples by sampling from prior $p_g(z)$. We effectively train G to minimize $\log(1-D(G(z)))$ and D to maximize $\log(D(x))$

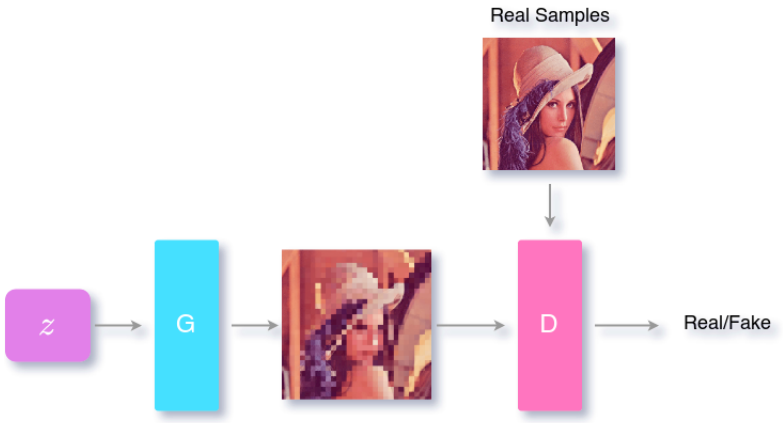

Figure 2.1.5: Illustration of GAN architecture. G is the generator network and D is the discriminator network.

making the loss function as $\mathcal{L}_{\text{GAN}}$ Eq. 2.5.

$$L_{\text{GAN}} = \mathbb{E}_{x \sim p_r(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.5}$$

However, when training GANs, practice is very different from theory. The training of the discriminator and generator is done iteratively and sequentially. But training the discriminator network to the optimal solution and then training the generator and repeating this loop is computationally challenging and would lead to overfitting the models on the finite dataset. To avoid this, the discriminator is trained for k mini-batch iterations before training the generator for one iteration. This is to keep the discriminator close to optimality while slowly training the generator [17]. The problem that arises here is that the generated samples are drastically different from the real samples as the generator has not yet learned to produce good samples. As the generator outputs samples close to noise, the discriminator easily distinguishes these samples from the real samples with high confidence. This saturates the loss term $\log(1 - D(G(z)))$ very quick and leads to the problem of *Vanishing Gradients*. Hence in practice, we train the generator G to maximize $\log D(G(z))$ instead of minimizing $\log(1 - D(G(z)))$ preventing the gradients from vanishing.

$$\begin{aligned} D^*_G(x) =& \frac{p_r(x)}{p_r(x) + p_g(x)} \\ =& \frac{1}{2} \end{aligned} \tag{2.6}$$

At global optimality $p_g = p_r$ and for a given generator G, the discriminator at optimality is $D^*_G(x)$ Eq. 2.6. Hence the virtual cost $C(G)$ [17] when $p_g = p_r$ is:

$$\begin{aligned} C(G) &= \max_D V(G, D) \\ &= \mathbb{E}_{x\sim p_r}\left[\log D^*_G(x)\right] + \mathbb{E}_{z\sim p_z}\left[\log\left(1 - D^*_G(G(z))\right)\right] \\ &= \mathbb{E}_{x\sim p_r}\left[\log D^*_G(x)\right] + \mathbb{E}_{x\sim p_g}\left[\log\left(1 - D^*_G(x)\right)\right] \\ &= \mathbb{E}_{x\sim p_r}\left[\log \frac{p_r(x)}{P_r(x) + p_g(x)}\right] + \mathbb{E}_{x\sim p_g}\left[\log \frac{p_g(x)}{p_r(x) + p_g(x)}\right] \\ &= -\log(4) + D_{\mathrm{KL}}\left(p_r \| \frac{p_r + p_g}{2}\right) + D_{\mathrm{KL}}\left(p_g \| \frac{p_r + p_g}{2}\right) \\ &= -\log(4) + 2 \cdot \mathrm{JSD}(p_r \| p_g) \end{aligned} \tag{2.7}$$

The **JSD!** (**JSD!**) between two distributions is always non-negative and will be equal to zero only when both the distributions are equal [25]. As we derived in Eq. 2.7 above, the best value of C i.e $-\log 4$ is possible only when $p_g = p_r$. It is hard to stabilize the GAN's minimax game [3]. It requires carefully tuned hyperparameters to maintain an equilibrium between the two players. Failing to find the proper balance between the networks leads to the problem of *Non-Convergence*, where the training oscillates and never converge. When the generator is not strong enough and learns to produce samples that fool the discriminator, it eventually would restrict itself to only learn to produce such samples. This problem is referred to as *Mode Collapse* [10]. There are many hacks as well as principled approaches that are formulated to handle these problems with considerable success [47].

2.1.5 WGAN

WGANs is a variant of GANs, where the second network is a critic that scores the samples on how real they look rather than a discriminator that predicts binary labels of 1 and 0 for real or fake. WGAN use 1-Wasserstein distance [1] or Earth Mover's Distance (EM distance) instead of the Jensen–Shannon Divergence (JS-divergence) used in the standard discriminator based GAN. Since the Wasserstein distance is non-evaluative, a modified version Eq. 2.8 of it is proposed as the loss function in [4]. Where f being the critic network parameterized by w while clipping the weights to

satisfy Lipschitz constraint.

$$L = \mathbb{E}_{x \sim P_r} [f_w(x)] - \mathbb{E}_{x \sim P_g} [f_w(x)] \tag{2.8}$$

The fundamental goal of GANs is to minimize the distribution between the real and the generated distribution. This could be measured used either of KLD, JS-divergence, EM distance, or Wasserstein distance, the main difference being their impact on the convergence of these distributions. The interesting feature of the Wasserstein distance is that it is continuous and differentiable. Using this distance the critic can train till optimality while having a reliable gradient throughout the training. Hence the critic in WGAN does not have the saturation and vanishing gradient problems that exist in standard GANs. Due to continuous and clean gradients, the training is significantly stable and less sensitive to hyperparameters and model architecture. With WGAN the mode collapse problem is also significantly reduced. When it comes to practice, the most important problem that hinders training GANs is that there is no correlation of the quality of the generated data say, images, and the loss function. However, WGAN tries to converge the distributions while lowering the generation loss and considerable relation between the loss and the quality of generations can be observed. WGAN with Gradient Penalty (WGAN-GP) [19] is an improved WGAN that uses gradient penalty to enforce Lipschitz constraint.

2.1.6 Hybrids - VAEGAN

While the VAEs learn the latent space of the data very efficiently, the generative capabilities are limited in comparison to GANs. In the case of image generations, VAEs usually generate blurry images. While well trained GAN learn to generate photorealistic images. Though the task of the discriminator in GANs is to only learn what is real and what is fake, it implicitly learns rich a similarity metric in order do so [30]. The idea of a VAE-GAN, illustrated in fig. 2.1.6, is to exploit this ability of GANs as a learning metric for VAEs and the ability of VAEs to learn dense latent representation of the data.

Taking the example of images again, the element-wise error is a very poor similarity metric as a small deviation in a high-level feature, like eyebrow or head rotation would lead to high error as the pixel displacement propagates through huge parts of the image.

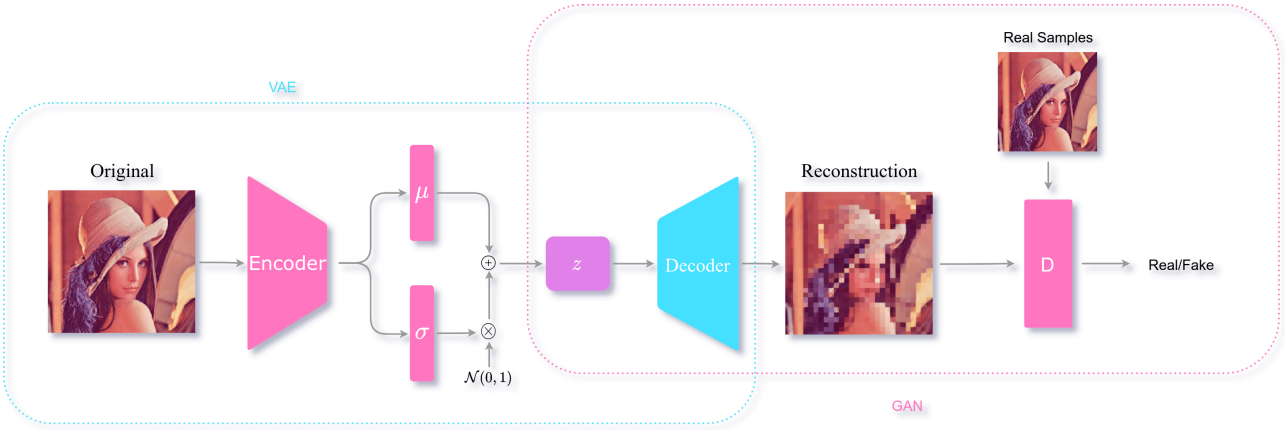

Figure 2.1.6: Illustration of the VAE-GAN architecture

These shifts, in reality, are plausible and realistic, probably indistinguishable for the human eye. Using the discriminator as a similar metric would address this problem as the error would be low for realistic deviations of features compared to unrealistic shifts say, the noise being upside down. This can be achieved by replacing the element-wise loss of the VAE Eq. 2.3 with the hidden representation $D_l(x)$ of an intermediate layer l in the discriminator that would correspond to the hidden similarity metric. The Gaussian distribution for $D_l(x)$ is :

$$p\left(D_l(x)|z\right) = \mathcal{N}\left(D_l(x)|D_l(\tilde{x}), I\right) \tag{2.9}$$

Where, $\tilde{x}$ is the generated sample from the VAE's decoder p. $D_l(tildex)$ is the mean of the Gaussian distribution and I is the identity covariance. Replacing this as the similarity metric in 2.3, we get the new $\mathcal{L}_{\text{recon}}$:

$$\mathcal{L}_{\text{recon}}^{D_l} = -\mathbb{E}_{q(z|x)} \log p\left(D_l(x)|z\right) \tag{2.10}$$

$\mathcal{L}_{\text{recon}}^{D_l}$ which uses the l^{th} layer of the discriminator is only the metric for the VAE, the VAE-GAN is trained on a triplet loss 2.11 with $\mathcal{L}_{GAN}$ from Eq. 2.5 as a *style error*. Here the generator model is the same as the decoder of the VAE as it maps from z to x just like G.

$$\mathcal{L}_{\text{VAEGAN}} = \mathcal{L}_{\text{recon}}^{D_l} + \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{GAN}} \tag{2.11}$$

Training GANs is hard but training VAE is harder. It is very important to consider

that the training of the VAE and the GAN takes place simultaneously. While doing so it is required to *limit the error propagation* of the triplet loss to the entire model. The discriminator should not learn to minimize $\mathcal{L}_{\text{recon}}^{D_l}$, if it does, the discriminator collapses. Better results are observed by restricting the error signal to reach the encoder q as illustrated in the fig. 2.1.7.

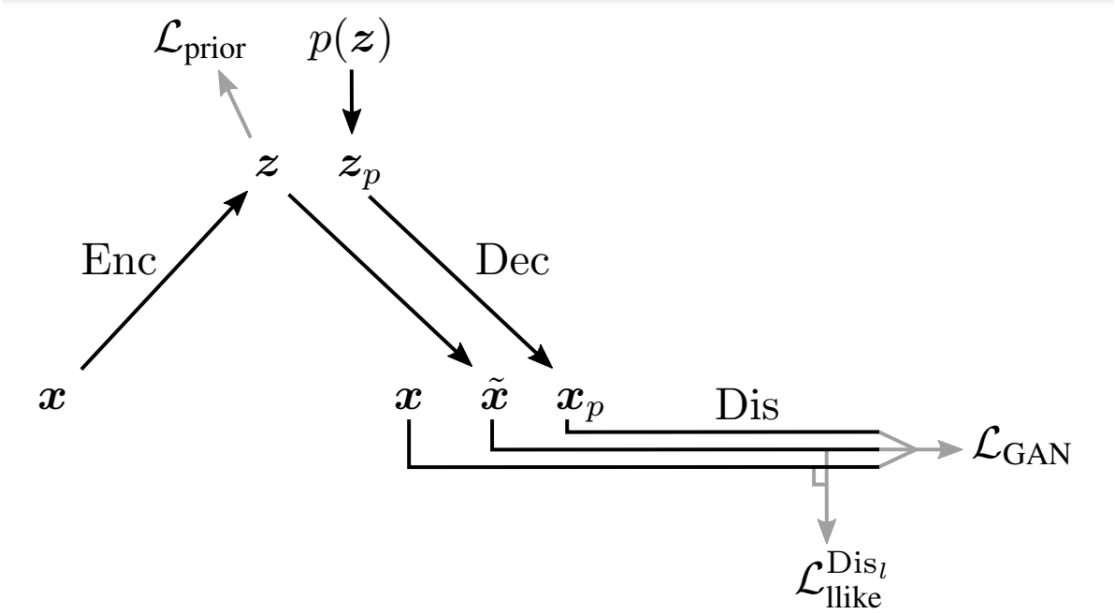

Figure 2.1.7: Illustration of the data flow and the loss of VAE-GAN, where $\mathcal{L}_{\text{recon}}^{D_l} = \mathcal{L}_{\text{llike}}^{Dis_l}$ Image Source [30]

.

As discussed in 2.1.3, VAE as a whole has two objectives - minimize the $\mathcal{L}_{\text{recon}}$ and the $\mathcal{L}_{\text{prior}}$ and a weighing factor β is used to maintain a trade-off betweent he qualilty of the reconstrution and the extent of disentanglement. Similarly, when it comes to VAE-GAN, the decoder alone has two objectives. One is to generate samples minimizing the $\mathcal{L}_{\text{recon}}^{D_l}$ and the other is to make sure that the generated samples can fool the discriminator. And the trade off is regulated by using γ to weigh $\mathcal{L}_{\text{recon}}^{D_l}$ and $\mathcal{L}_{\text{GAN}}$ as in Eq. 2.12.

$$\theta_p \overset{+}{t} -\nabla_{\theta_p} \left(\gamma L_{\text{uike}}^{D} - \mathcal{L}_{\text{GAN}}\right) \tag{2.12}$$

In the standard GAN training, samples from the prior $p(z)$ are passed to the decoder which generates samples that are then passed to the discriminator. Interesting observation when using VAE-GAN is, sampling x from the encoder $q(z|x)$ further improves the results. As the VAE tries to minimize $\mathcal{L}_{\text{prior}}$, the samples from $p(z)$ and

$q(z|x)$ while become similar during the training. As the generated samples $p(q(z|x))$ using the encoder are more realistic than $p(p_{prior}(z))$ using the prior, they serve as better adversarial examples for the discriminator. The $\mathcal{L}_{\text{GAN}}$ loss to be used to leverage this benefit is:

$$\mathcal{L}_{\text{GAN}} = \log(Dis(x)) + \log(1 - Dis(Dec(z))) + \log(1 - Dis(Dec(Enc(x)))) \tag{2.13}$$

2.2 Research Area Introduction

In this section, the details of various approaches and the SOTA in 3D HPE are presented along with some works in the sibling task of hand pose estimation. This section aims to give an overview of how the problem of 3D HPE is tackled in the literature.

2.2.1 Pose from images

Numerous works try to estimate 3D human poses from 2D RGB images or 2D joint confidence heatmaps [7, 9, 35, 39, 45]. Most of these methods follow a cascading approach, where an explicit intermediate representation of 2D pose or 2D heatmaps is used.

For example, [35] proposes a general framework with 3 networks. Human detection Network, RootNet, PoseNet. Where the human detection network predicts the region the human is in an image. The RootNet localizes the human's root in the global 3D world. And, the PoseNet predicts the 3D pose of a single person relative to the root. Where the root is a fixed reference point of the human body say, pelvis.

The advantage of such top-down frameworks is to divide the task of RGB to 3D into smaller, well-studied sub-tasks. This makes scaling single-person pose estimation algorithms for multi-person pose estimation easy, as the majority of the data available mostly consists of a single person per frame. In addition to it, this approach provides the opportunity to improve certain modules without affecting or having to re-train the other modules of the system.

2.2.2 Pose Lifting

In contrast to the estimating pose from an image, Pose Lifting works such as [7, 8, 29, 36, 44], focus on estimating 3D poses from 2D poses alone. Assuming 2D poses from the SOTA methods in 2D HPE. These methods include simple linear models as first described in [34] with a series of fully connected linear layers, and sometimes batch normalization, dropout and, residual connections to regress 3D pose effectively.

Non-Rigid Structure from Motion (NRSfM) is another promising lifting method that also leverages images along with 2D annotations. NRSfM deals with the problem of reconstructing 3D shape (pose/point cloud) and cameras of each projection from a

sequence of images with corresponding 2D orthogonal projections (2D keypoints). This approach has been widely used in facial keypoint detection and [28] introduces deep learning variant for the same. Instead of predicting the 3D coordinates of each keypoint/joint of the 3D pose, [28, 36, 45, 46] predicts the 3D shape and camera pose from 2D pose using this method.

The Pose Lifting approach facilitates to leverage the already well established 2D HPE models that are trained on enormous and diverse labeled data. Thus demanding lesser training data for 3D pose estimation than it would need when learning from images. Since these networks do not have large convolution layers they are less computationally inexpensive for both training and inference on edge computing units. Moreover, the 2D and 3D pose data usually can be entirely loaded onto the GPU further accelerating the training procedure. Thus addressing the critical problem that hinders scalability of 3D HPE models and also helps to develop better modular systems by combining the best of Pose Lifting networks with the best of 2D HPE. However, due to the inherent ambiguity in lifting pose to 3D and as the images are not captured with orthogonal cameras, reprojection of 3D pose will vary from the ground truth, it is challenging to match the performance of models trained on 3D ground truth.

2.2.3 Non-Supervised Learning

The standard way to train 3D/2D HPE is by minimizing the distance between the predicted 3D/2D pose and its corresponding ground truth. The area of 2D HPE is well established and matured with reliable systems deployed in the real world. This was made possible with the high volume of images from diverse settings and the reasonable ease of manual labeling of 2D poses. On the other hand, labeling 3D pose manually is not practical. Though single-person datasets such as Human3.6M [24], Human Eva [23] and, multi-person datasets such as CMU Panoptic [20] provide 3D pose ground truth. They are obtained using Motion Capture (MoCap) systems which are only limited to indoors or cannot be directly adapted to outdoor environments where the majority of the use cases exist fig[2.2.1]. It is also worth mentioning JTA(Joint Track Auto) dataset [14] that is made using the GTA(Grand Theft Auto) game engine which is technically scalable with its own limitations. But datasets from simulations come with the difficulty of domain adaptation to be transferable to the real world.

To overcome this bottleneck, [29] proposes unsupervised training of a generative

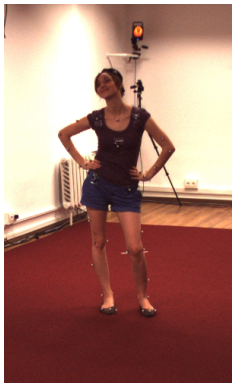

Figure 2.2.1: Image from Human3.6 Dataset [24] of subject wearing MoCap markers

adversarial network by projecting the predicted 3D pose back to 2D and minimizing its distance with the input 2D pose. And further training a discriminator to distinguish the real 2D pose from the projected poses. Thus removing the need for any explicit 3D annotations besides 2D pose that are either manually labeled or obtained using 2D HPE models. RepNet [44] trains an adversarial network without 2D-3D correspondences in a weakly supervised manner. Moreover, it also does not require camera parameters to project the 3D pose but learns to predict them. Thus enabling better generalization to more diverse data with unknown cameras and poses.

To test the maximum capability of Pose Lifting networks, [8] proposes a combination of unsupervised and adversarial learning that mainly leverages the property of *plane-invariance*. It is the property that 2D projections of a 3D pose from different camera viewpoints, when lifted should produce identical and the original 3D pose. In this method, the predicted 3D pose is rotated in random angles and is reprojected to 2D in a different Point of View (POV). A discriminator is then used to evaluate if this new 2D pose is in the possible pose distribution which is learned from 2D pose datasets alone. These steps are redone in reverse order to obtain the original 2D input. This cycle provides three intermediate representations of the single 2D input that the models learn from. Additionally, this approach exploits the temporal consistency in the datasets as well as integrates a domain adaptation network to learn from different datasets and distributions to achieve comparable results to that of the methods that require more supervision.

2.2.4 Multimodal Representation Learning

Another interesting approach is training VAEs using multiple modalities like images, poses, depth maps [18, 40, 41, 43]. Multimodal Variational Auto-Enocder (MVAE)s learn representation from different modalities in the same latent space. True multimodal learning needs to fulfill 4 criteria as follows: i) *Latent Factorization* - Implicit factorization of latent space into private, shared subspaces based on modality as illustrated in the figure[2.2.2]. ii) *Coherent Joint Generation* - Coherence in generations of different modalities from the same latent value with respect to the shared aspects of the latent. iii) *Coherent Cross Generation* - Generation of one modality conditioned on data from different modality while preserving the similarity between them. iv) *Synergy* Enhancement in generation quality of one modality as a result of learning representations of different modalities.

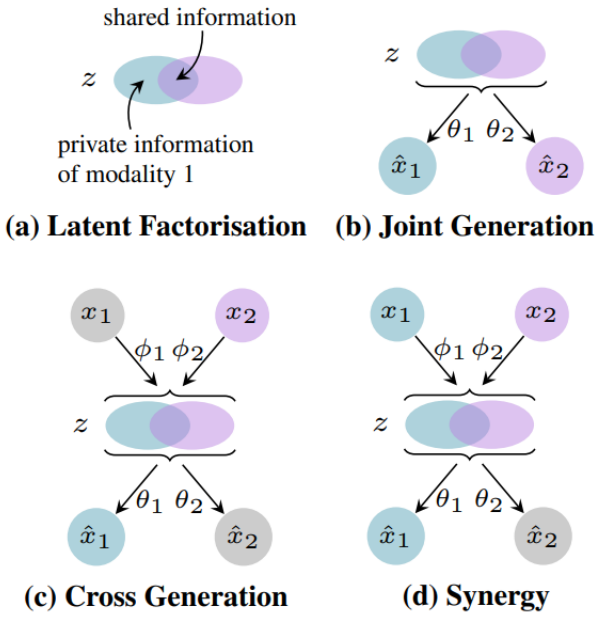

Figure 2.2.2: Criteria for Multimodal Generation [40]

Mixture-of-Experts Multimodal Variational Auto-Enocder (MMVAE) proposed by [40] fulfills all 4 of the above-mentioned criteria learning representations of image and text data, while other approaches focus on leveraging specific advantages of multimodal learning. Consider the cross-modal learning for 3D Hand Pose Estimation proposed by [41]. It involves training an encoder-decoder pair to learn image representation, and another such pair to learn 3D hand pose representations in the same latent space. This training procedure focuses on cross-generation and synergy. That is, using the shared latent space of the image and pose representations, the RGB

image encoder combined with the pose decoder can generate 3D poses and vice versa while preserving the commonality between the conditioned and the generated data. With this approach, it is possible to train a VAE for 3D HPE from RGB images without explicit intermediate stages like the earlier mentioned cascading approaches. Making it more efficient and fast for both training and inference without compromising the modularity offered by cascading approaches.

2.3 Related Work - A closer look

In this section, works that are directly related to the thesis are discussed in more detail. Some are the best examples of their kind and have already been discussed thoroughly. The basic idea of the thesis is to learn 3D HPE just from 2D pose data without using 3D ground truth in any shape or form. Thus developing a method that can exploit the huge amounts of 2D pose data that can be generated using state of the art 2D pose networks on diverse images from the real world. The following approaches use weakly supervised or unsupervised approaches to accomplish the same. These serve as the inspiration for many of the choices taken in this thesis and also help understand the possibilities of reducing the need for explicit 3D supervision.

To the best of knowledge acquired during the period of the thesis, [8, 13, 29, 36] are the main approaches that do not use 3D supervision in any way. While [32, 44] are among the main approaches that use 3D supervision to train the discriminator alone. The approaches that are not mentioned are either the approaches the above mentioned are built up or have been missed during the literature study.

[8, 13, 29] can be viewed as a series of approaches that are built on one another in the same order. They take 2D poses as the input and learn to predict the depth offset for each joint to reconstruct 3D. Out of the three Ching et al. [8], using the plane invariance, geometric self-supervision, and adversarial learning as discussed in 2.2, achieves the SOTA results compared to fully supervised methods and also present ways to use domain adaptation network, temporal consistency to further integrate more datasets and improve the performance. Thus directly address the hurdles of scaling the 3D HPE network to the real world. However, they also acknowledge the fact that most of the predictions made by SOTA 2D HPE model on real-world images had missing joints. Since the proposed approaches only predict the depth of every joint, the error from the 2D input pose is directly propagated to the 3D prediction. More importantly,

it is not possible to use most of the data that is generated from 2D pose models. Hence it is very crucial to handle the problem of ***missing joints*** to truly unlock the potential of unsupervised learning.

Wandt et al. [44] also discussed in 2.2 proposes an architecture that learns to predict the whole 3D pose, while also learning the camera parameters that are used to project the predict 3D to 2D. The idea behind the camera network is to learn the view angle given pose to generalize to unknown cameras. The pose network learns to converge the predicted 2D reprojections, while using a GAN trained on ***3D ground truth labels*** to supervise the predicted 3D pose. Though there is no direct error propagation from 2D input to 3D, it is important to note the problem of missing joints is not yet addressed.

However, there another fundamental problem persists. As mentioned in 1.1, 2D-to-3D pose lifting is an ill-posed-inversed problem due to ***depth ambiguity*** as there are multiple plausible 3D poses that gives the same 2D projection. Addressing this problem, Chen Li et al. [31] proposes a variational inference model inspired by the architecture of Wandt et al. [44] that takes 2D pose along with a *latent code* to produce a 3D pose. The predicted 3D pose varies according to the latent code while maintaining the same 2D reprojection. This variational inference of 3D pose addresses both the problems of depth ambiguity and missing joints.

The above approach has been further improved by Chen Li et al. in [32] to leverage adversarial training using GAN trained using 3D ground truth. Though there is no direct supervision on the 3D pose, 3D data is still required to train this network. It is important to note that in the case where a large amount of 3D ground truth poses are available, they can be easily exploited to generate 2D poses of large volume and it is not practical to get 3D poses in the wild to scale the models. In this thesis, we try to address all the aforementioned problems.

Chapter 3

Data

This chapter discusses the datasets used in the thesis, as well as the processing steps to make the data more learnable. The main dataset used in the thesis is *Human3.6M*: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments [24]. Most of the related works benchmark their methods on Human3.6M and it also is freely accessible to academics on request. For further evaluation of model performance in the wild, outdoor datasets that do not have 3D ground truth such as *3DPW*: 3D Poses in the Wild [33] would be used.

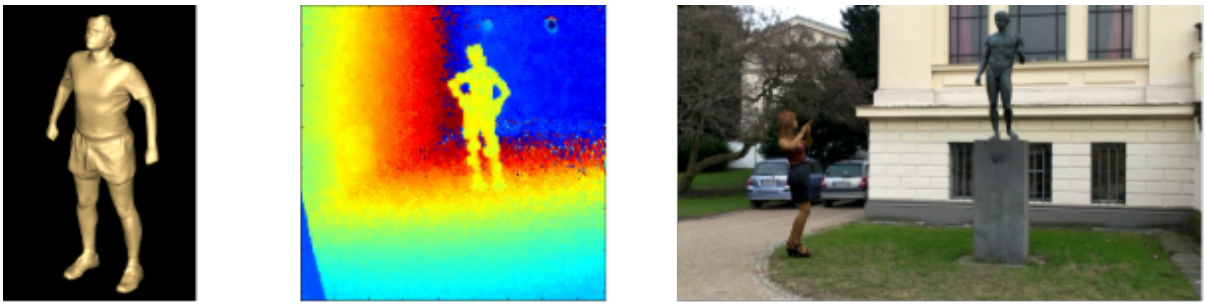

Figure 3.0.1: Full body model, depth from time of flight and mixed reality in Human3.6M dataset

3.1 Human3.6M

Human3.6M is a large scale indoor dataset with 3.6 million human poses collected with 4 cameras at different angles using a highly accurate maker-based MoCap system. The dataset constitutes 15 diverse motion and actions such as eating, sitting, walking in various everyday scenarios such as a hand in the pocket, talking over the phone,

walking a dog, etc. These actions are performed by 11 professional actors wearing a variety of realistic clothing. The datasets provides synchronised 2D and 3D data including full-body scans as shown in figure[3.0.1]. It also includes mixed-reality test data created using animated human models to cover huge variations of background, clothing, illumination, occlusion, and camera angles.

3.1.1 Depth Ambiguity and Camera Modeling

The projection of the poses in the thesis is using a simple pinhole camera model as illustrated in fig.3.1.1. A unit camera model is assumed for the dataset and the poses are predicted around the origin and translated to a fixed image plane. Thus the error in projection is not significant.

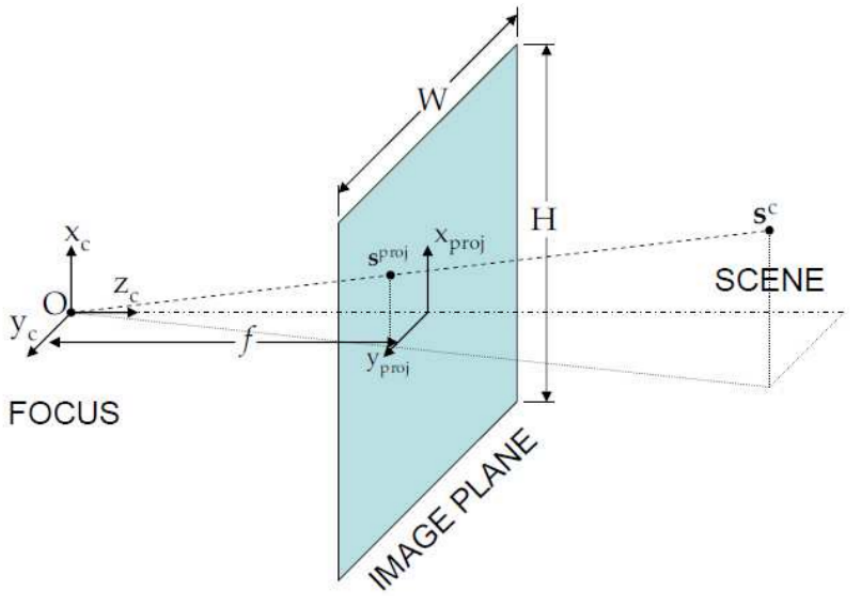

Figure 3.1.1: Pinhole Camera Model. Image Source [2]

One of the main problems discussed in 2.3 is depth ambiguity. The fig 3.1.2 illustrates the challenges in lifting 2D pose to 3D pose. During evaluation under protocol 2, the 3D poses are transformed using rigid or Procrustes alignment with the ground truth pose as illustrated in fig. 3.2.1. But in the proposed method we only translated and rotate but do not scale the pose.

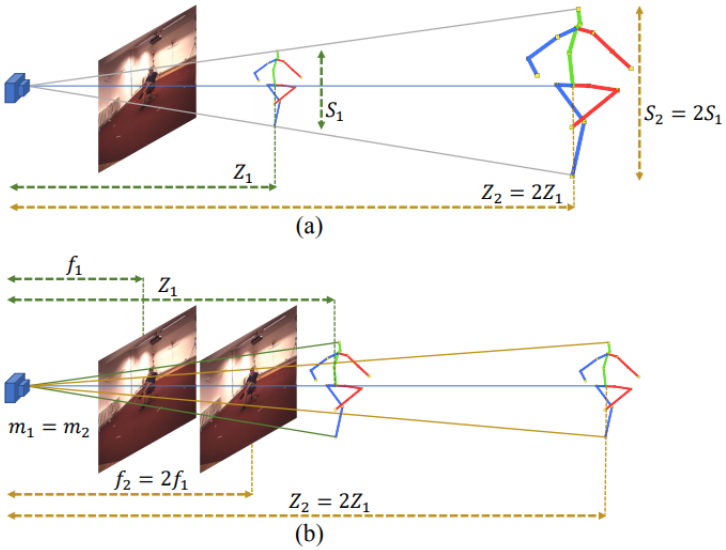

Figure 3.1.2: Illustration of depth ambiguity. (a) shows 2 of the infinite possible poses that result in the same 2D reprojection. Where (b) shows the same phenomenon for different focal legths Image Source [7]

3.2 Processing

The methods explored by this thesis would require only images, 2D, and 3D human pose from the dataset. The following are the pre-processing steps for the 2D and 3D poses.

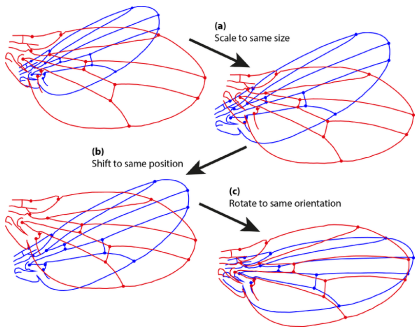

Figure 3.2.1: Illustration of Procrustes Alignment. Image Source [37]

The 3D pose in the dataset that is obtained from the marker-based MoCap is in a global reference frame. These poses using the camera parameters are transformed into the camera coordinate frame. For the task of predicting 3D pose from either images or 2D pose, it is unrealistic to directly estimate all the joints of the pose in a global frame. So the first step of processing would be to zero the pose w.r.t the root joint say, Pelvis. As the root is always zero, we remove it so we do not have to learn the constant joint. Removing Pelvis, 16 out of the 17 joints or keypoints remain. The 3D pose is further

scaled in down so that the distance between the root and the head is 1. This results in numerical stability during training.

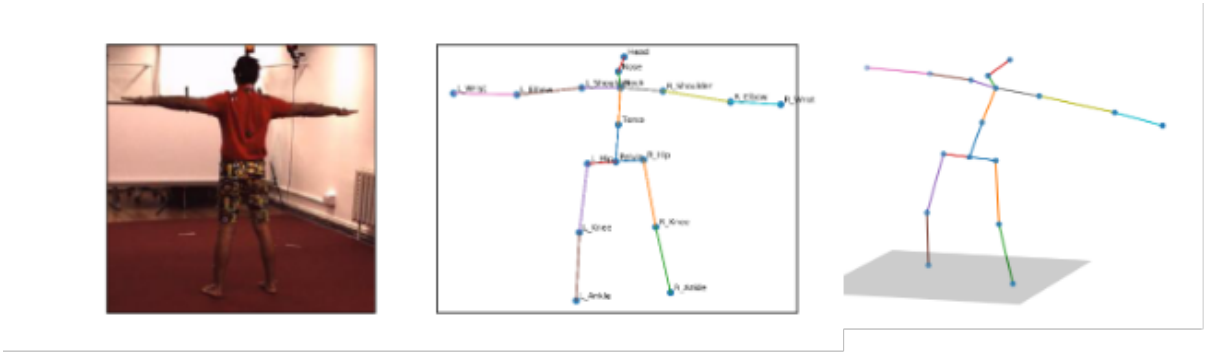

Figure 3.2.2: Human3.6M Pose Sample

The 2D pose which is obtained from the 3D pose, is also in the camera coordinate frame. The 2D pose is also zeroed and scaled so that the distance between the root and head is 1/c units. Where c is a constant distance at which image place is fixed. As a unit camera or camera with unit focal length is assumed the projection of unit 3D pose onto a plane at a fixed distance 1/c units. The root of the 2D pose is however removed to remain consistent with the 3D pose. An image sample from the dataset with its corresponding 2D and 3D pose is illustrated in the figure[3.2.2].

The estimated poses from the networks that are trained on downscaled poses are upscaled to the original size using for valuation. This postprocessing step is required for getting the distance between prediction and ground truth keypoints in true units of millimeters.

Chapter 4

Method

The initial method that is being explored is using VAE to estimate 3D pose from both RGB image and 2D pose. This approach is based on the crossmodal hand pose estimation [41] but with the goal to test the performance on human poses and to investigate other ideas and techniques that the paper has not addressed. Currently we explore the image and pose modalities and investigate training VAE for image to 3D and 2D to 3D pose estimation.

4.1 Architecture

4.1.1 VAE

As described in section[2.2.4], the crossmodal training involves training the encoders of each modality to learn to represent the input in the same latent space. Similarly the decoders learn to sample an embedding from this shared latent space and reconstruct an image or pose respectively. In contrast to the [41], that uses an RGB to RGB and 3D to 3D encoder-decoder pair to make enable self-supervision, we use 2D encoder instead of a 3D to evaluate cross-generation and synergy for 3D HPE from images or 2D pose. The prediction of the 3D decoder could be reprojected to 2D to eliminate the need for 3D annotation.

4.1.2 Discriminator

To leverage the power of transfer learning, a pre-trained ResNet-18 [21] with two additional linear layers one for mean and another for log-variance is used as the encoder and a series of five 2D convolutional layers, each followed by a batch normalization and an activation function like ReLU or Tahn is used as the image decoder.

4.1.3 Hybrid

For the sake of simplicity and consistency with the previous works for benchmarking the performance, we use a series of 5 linear and ReLU activation blocks with additional linear layers for mean and log-variance for the 2D pose encoder and a linear later for upsampling means to hidden dimensions of the main linear blocks.

4.2 Training Scheme and Loss Function

Training a VAE is a notoriously difficult task, as it involves optimizing not just the reconstruction loss but also the KLD loss. With crossmodal training the number of metrics to optimize increases multi-fold. As described in section[2.2.4], The training scheme for crossmodal training (for crossmodal generation) involves training combinations of encoder and decoder of either the same or different modalities in the same epoch. The reconstruction loss function of that particular combination depends on the decoder. The image decoder uses L1 loss and the 3D pose decoder uses MSE loss. Though the KLD loss is the same for both, it is normalized with the number of elements in the reconstruction, i.e 16*3 for 3D pose and 256*256*3 for RGB images.

4.3 Bag of tricks

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat.

Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui

cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium

ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi.

4.4 Evaluation Metrics

3D human pose and Human3.6M in particular is mainly evaluated by Mean Per Joint Position Estimate (MPJPE) metric. MPJPE as it literally abbreviates, is the mean of the position estimate for all the joints of a pose. Where per-joint position estimate is nothing but the euclidian distance (usually measured in mm) between the predicted joint to its ground truth.

Chapter 5

Results

5.1 Human3.6M

5.1.1 Evaluation protocol

Human3.6M has 11 subjects out of which 7 are publically released while the rest are kept private. There are 2 widely used evaluation protocol. Protocol-1 is using all 4 camera views in subjects $S1$, $S5$, $S6$, $S7$ and $S8$ for training and the same 4 camera views in subjects $S9$ and $S11$ for validation/testing. Protocol-2 is the same as 1 expect that the predictions are post-processed via a rigid transformation before comparing to the ground-truth.

5.1.2 2D-3D lifting

The current experiments and results are only for 2D to 3D lifting VAE and could be found at https://app.wandb.ai/b-sridatta/hpe3d?workspace=user-b-sridatta. The results illustrated below are after training the model for 100 epochs on around 400,000 2D poses without augmentation. The architecure is as described in earlier with 512 hidden units per linear layer and 100 latent dimension and with a beta weight for KLD loss as 0.001. Further immediate experiments that have to be carried out are itegrating data augmentation, beta annealing, lower latent dimensions and image modalities.

Figure [5.1.1] illustrates 25 random predictions of the validation poses (in blue) and their correspsonding error (in red) w.r.t the ground truth (in gray) in millimeters.

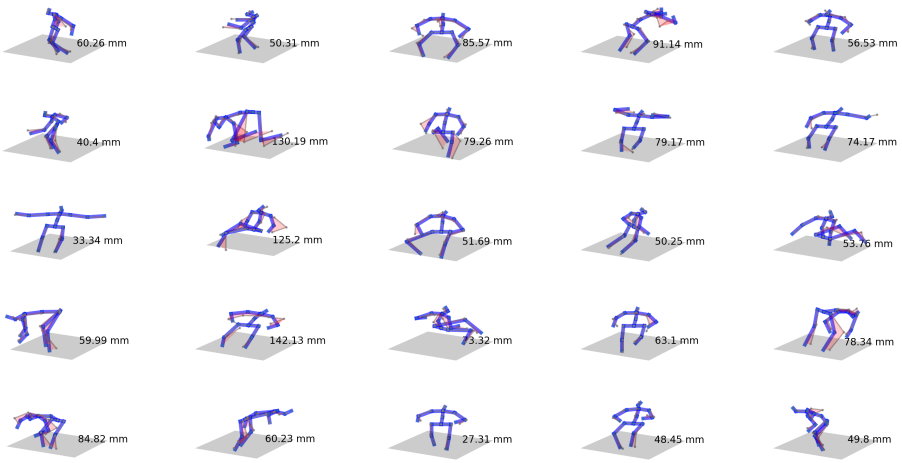

Figure 5.1.1: Comparing predictions and ground truth

Figure[5.1.2] is the visualization of 2D pose embedding in latent space after dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) with different number of neighbours where small neighbours extract local features and large numbers extract more global features. Each action is given a unqiue color. Though we small clusters of blues, browns, pinks the overall space looks very mixed up. This is expected to be improved after annealing beta from 0 to 1 over t he course of training or by using cyclical annealing [16]. However, it could also be the case that the many of the instance in different actions over lap. For example, the action standing up and sitting down have instances while both or standing or sitting etc. This could be verified by visualising the latent space with images rather than just points.

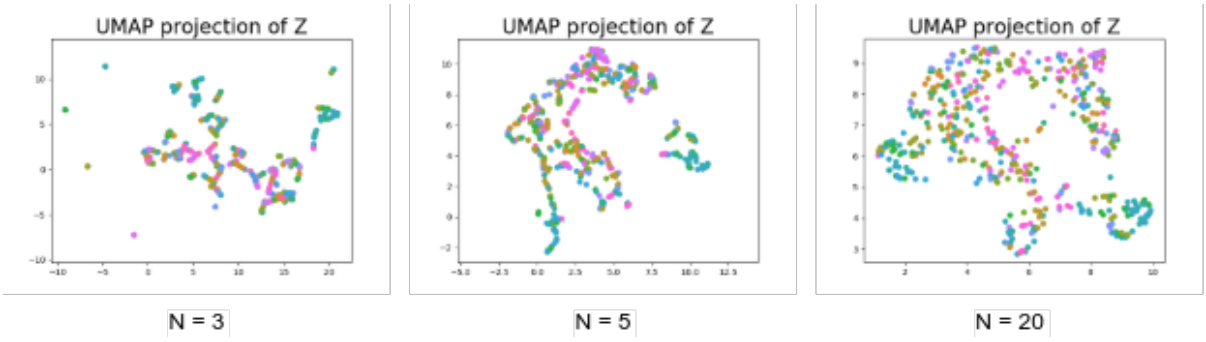

Figure 5.1.2: UMAP Visualization of samples in latent space with varying nearest neigbours

Chapter 6

The work

Describe the degree project. What did you actually do? This is the practical description of how the method was applied.

6.1 Implementation

6.1.1 Developments

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

6.1.2 Monitoring

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu

libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Chapter 7

<Conclusions>

Describe the conclusions (reflect on the whole introduction given in Chapter 1).

Discuss the positive effects and the drawbacks.

Describe the evaluation of the results of the degree project.

Describe valid future work.

The sections below are optional but could be added here.

7.1 Discussion

7.1.1 Future Work

7.1.2 Final Words

Bibliography

[1] 2020. URL: https://en.wikipedia.org/wiki/Wasserstein_metric.

[2] *(10) (PDF) Fusion of Imaging and Inertial Sensors for Navigation*. https://www.researchgate.net/publication/35152505_Fusion_of_Imaging_and_Inertial_Sensors_for_Navigation. (Accessed on 09/26/2020).

[3] Arjovsky, Martin and Bottou, Léon. *Towards Principled Methods for Training Generative Adversarial Networks*. 2017. arXiv: 1701.04862 [stat.ML].

[4] Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. *Wasserstein GAN*. 2017. arXiv: 1701.07875 [stat.ML].

[5] Ballard, Dana H. "Modular Learning in Neural Networks." In: *AAAI*. 1987, pp. 279–284.

[6] Blei, David M., Kucukelbir, Alp, and McAuliffe, Jon D. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (Apr. 2017), pp. 859–877. ISSN: 1537-274X. DOI: 10.1080/01621459.2017.1285773. URL: http://dx.doi.org/10.1080/01621459.2017.1285773.

[7] Chang, Ju Yong, Moon, Gyeongsik, and Lee, Kyoung Mu. "AbsPoseLifter: Absolute 3D Human Pose Lifting Network from a Single Noisy 2D Human Pose". In: *arXiv preprint arXiv:1910.12029* (2019).

[8] Chen, Ching-Hang, Tyagi, Ambrish, Agrawal, Amit, Drover, Dylan, Rohith, M. V., Stojanov, Stefan, and Rehg, James M. "Unsupervised 3D Pose Estimation with Geometric Self-Supervision". In: *CoRR* abs/1904.04812 (2019). arXiv: 1904.04812. URL: http://arxiv.org/abs/1904.04812.

[9] Cheng, Yu, Yang, Bo, Wang, Bo, Wending, Yan, and Tan, Robby T. "Occlusion-Aware Networks for 3D Human Pose Estimation in Video". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 723–732. DOI: 10.1109/ICCV.2019.00081. URL: https://doi.org/10.1109/ICCV.2019.00081.

[10] *Common Problems Generative Adversarial Networks Google Developers*. URL: https://developers.google.com/machine-learning/gan/problems.

[11] Contributors to Wikimedia projects. *Artificial neural network - Wikipedia*. [Online; accessed 11. Sep. 2020]. 2020. URL: https://en.wikipedia.org/w/index.php?title=Artificial_neural_network&oldid=975746760.

[12] Dario Amodei, Danny Hernandez. *AI and Compute*. https://openai.com/blog/ai-and-compute/. (Accessed on 05/17/2020).

[13] Drover, Dylan, MV, Rohith, Chen, Ching-Hang, Agrawal, Amit, Tyagi, Ambrish, and Huynh, Cong Phuoc. *Can 3D Pose be Learned from 2D Projections Alone?* 2018. arXiv: 1808.07182 [cs.CV].

[14] Fabbri, Matteo, Lanzi, Fabio, Calderara, Simone, Palazzi, Andrea, Vezzani, Roberto, and Cucchiara, Rita. "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World". In: *European Conference on Computer Vision (ECCV)*. 2018.

[15] Fan, Jianqing, Ma, Cong, and Zhong, Yiqiao. *A Selective Overview of Deep Learning*. 2019. arXiv: 1904.05526 [stat.ML].

[16] Fu, Hao. "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing". In: *NAACL*. 2019.

[17] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[18] Gu, Jiajun, Wang, Zhiyong, Ouyang, Wanli, Zhang, Weichen, Li, Jiafeng, and Zhuo, Li. "3D Hand Pose Estimation with Disentangled Cross-Modal Latent Space". In: *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 2020, pp. 380–389. DOI: 10.1109/WACV45572.2020.9093316. URL: https://doi.org/10.1109/WACV45572.2020.9093316.

[19] Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG].

[20] Hanbyul Joo, Hao Liu. "Panoptic Studio: A Massively Multiview System for Social Motion Capture". In: (2015).

[21] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[22] Hinton, G. E. "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786 (2006), pp. 504–507. DOI: 10.1126/science.1127647. URL: http://dx.doi.org/10.1126/science.1127647.

[23] *HumanEva : Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion | SpringerLink*. https://link.springer.com/article/10.1007/s11263-009-0273-6. (Accessed on 06/08/2020).

[24] Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1325–1339.

[25] *Jensen–Shannon divergence*. Sept. 2020. URL: https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence.

[26] Jordan, Jeremy. *Variational autoencoders*. July 2018. URL: https://www.jeremyjordan.me/variational-autoencoders/.

[27] Kingma, Diederik P and Welling, Max. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML].

[28] Kong, Chen and Lucey, Simon. "Deep Non-Rigid Structure From Motion". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1558–1567. DOI: 10.1109/ICCV.2019.00164. URL: https://doi.org/10.1109/ICCV.2019.00164.

[29] Kudo, Yasunori, Ogaki, Keisuke, Matsui, Yusuke, and Odagiri, Yuri. “Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations”. In: *CoRR* abs/1803.08244 (2018). arXiv: 1803.08244. URL: http://arxiv.org/abs/1803.08244.

[30] Larsen, Anders Boesen Lindbo, Sønderby, Søren Kaae, Larochelle, Hugo, and Winther, Ole. *Autoencoding beyond pixels using a learned similarity metric*. 2015. arXiv: 1512.09300 [cs.LG].

[31] Li, Chen and Lee, Gim Hee. *Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network*. 2019. arXiv: 1904.05547 [cs.CV].

[32] Li, Chen and Lee, Gim Hee. *Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses*. 2020. arXiv: 2008.05770 [cs.CV].

[33] Marcard, Timo von, Henschel, Roberto, Black, Michael, Rosenhahn, Bodo, and Pons-Moll, Gerard. “Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera”. In: *European Conference on Computer Vision (ECCV)*. 2018.

[34] Martinez, Julieta, Hossain, Rayat, Romero, Javier, and Little, James J. “A Simple Yet Effective Baseline for 3d Human Pose Estimation”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 2659–2668. DOI: 10.1109/ICCV.2017.288. URL: https://doi.org/10.1109/ICCV.2017.288.

[35] Moon, Gyeongsik, Chang, Ju Yong, and Lee, Kyoung Mu. “Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image”. In: *CoRR* abs/1907.11346 (2019). arXiv: 1907.11346. URL: http://arxiv.org/abs/1907.11346.

[36] Novotny, David, Ravi, Nikhila, Graham, Benjamin, Neverova, Natalia, and Vedaldi, Andrea. “C3DPO: Canonical 3d pose networks for non-rigid structure from motion”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7688–7697.

[37] *Procrustes analysis - Wikipedia*. https://en.wikipedia.org/wiki/Procrustes_analysis. (Accessed on 09/26/2020).

[38] Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[39] Sharma, Saurabh, Varigonda, Pavan Teja, Bindal, Prashast, Sharma, Abhishek, and Jain, Arjun. *Monocular 3D Human Pose Estimation by Generation and Ordinal Ranking*. 2019. arXiv: 1904.01324 [cs.CV].

[40] Shi, Yuge, Siddharth, N., Paige, Brooks, and Torr, Philip H. S. "Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models". In: *CoRR* abs/1911.03393 (2019). arXiv: 1911.03393. URL: http://arxiv.org/abs/1911.03393.

[41] Spurr, Adrian, Song, Jie, Park, Seonwook, and Hilliges, Otmar. "Cross-Modal Deep Variational Hand Pose Estimation". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 89–98. DOI: 10.1109/CVPR.2018.00017. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Spurr%5C_Cross-Modal%5C_Deep%5C_Variational%5C_CVPR%5C_2018%5C_paper.html.

[42] Vincent, Pascal, Larochelle, Hugo, Bengio, Yoshua, and Manzagol, Pierre-Antoine. "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1096–1103.

[43] Wan, Chengde, Probst, Thomas, Gool, Luc Van, and Yao, Angela. "Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1196–1205. DOI: 10.1109/CVPR.2017.132. URL: https://doi.org/10.1109/CVPR.2017.132.

[44] Wandt, Bastian and Rosenhahn, Bodo. "RepNet: Weakly Supervised Training of an Adversarial Reprojection Network for 3D Human Pose Estimation". In: *CoRR* abs/1902.09868 (2019). arXiv: 1902.09868. URL: http://arxiv.org/abs/1902.09868.

[45] Wang, Chaoyang, Kong, Chen, and Lucey, Simon. “Distill Knowledge From NRSfM for Weakly Supervised 3D Pose Learning”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 743–752. DOI: 10.1109/ICCV.2019.00083. URL: https://doi.org/10.1109/ICCV.2019.00083.

[46] Wang, Chaoyang, Lin, Chen-Hsuan, and Lucey, Simon. “Deep NRSfM++: Towards 3D Reconstruction in the Wild”. In: *CoRR* abs/2001.10090 (2020). arXiv: 2001.10090. URL: https://arxiv.org/abs/2001.10090.

[47] Weng, Lilian. *From GAN to WGAN*. 2019. arXiv: 1904.08994 [cs.LG].

[48] Xie, Junyuan, Xu, Linli, and Chen, Enhong. “Image denoising and inpainting with deep neural networks”. In: *Advances in neural information processing systems*. 2012, pp. 341–349.

[49] Zhang, Richard, Isola, Phillip, and Efros, Alexei A. “Colorful image colorization”. In: *European conference on computer vision*. Springer. 2016, pp. 649–666.