



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

Unsupervised 3D Human Pose Estimation

**KTH Thesis Report
Draft for final thesis meeting**

Sri Datta Budaraju

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

****YET TO BE WRITTEN****

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 1/2 A4-page. English abstract

Keywords

Template, Thesis, Keywords ...

Abstract

YET 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 lösas i materialet. Presentera

Nyckelord

Kandidat examensarbete, ...

Acknowledgements

****YET TO BE WRITTEN****

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
- β -VAE** Beta Variational Auto-Encoder
- EM distance** Earth Mover's Distance
- FC** Fully Connected
- GAN** Generative Adversarial Network
- HPE** Human Pose Estimation
- JS-divergence** Jensen–Shannon Divergence
- KLD** Kullback–Leibler Divergence
- L1** Least Absolute Deviations
- MMVAE** Mixture-of-Experts Multimodal Variational Auto-Encoder
- MoCap** Motion Capture
- MPJPE** Mean Per Joint Position Estimate
- MSE** Mean Squared Error
- MVAE** Multimodal Variational Auto-Encoder
- NRSfM** Non-Rigid Structure from Motion
- PJPE** Per Joint Position Estimate
- POV** Point of View
- ReLU** Rectified Linear Unit
- SOTA** State-of-The-Art
- UMAP** Uniform Manifold Approximation and Projection
- VAE** Variational Auto-Encoder
- WGAN-GP** Wasserstein Generative Adversarial Network (WGAN) with Gradient Penalty
- WGAN** Wasserstein Generative Adversarial Network
- GMM** Gaussian Mixture Model

MDN Mixture Density Network

Contents

1	Introduction	1
1.1	Background	1
1.2	Problem	2
1.3	Goal	2
1.4	Benefits, Ethics and Sustainability	3
1.5	Methodology	3
1.6	Stakeholders	4
1.7	Delimitations	4
1.8	Outline	5
2	Theoretical Background	6
2.1	Preliminary Concepts	6
2.1.1	Autoencoder	6
2.1.2	Variational Autoencoders	7
2.1.3	Beta Variational Autoencoder	9
2.1.4	Generative Adversarial Networks	10
2.1.5	WGAN	12
2.1.6	Hybrids - VAEGAN	13
2.2	Research Area Introduction	17
2.2.1	Cascading Approach	17
2.2.2	Pose Lifting	18
2.2.3	Multiple Hypothesis Estimation	19
2.2.4	Non-Supervised Learning	20
2.2.5	Multimodal Representation Learning	22
2.3	Related Work	24
3	Data	27

CONTENTS

3.1 Human3.6M	28
3.2 Projective Geometry	29
3.3 Depth Ambiguity and Camera Modeling	29
3.4 Processing	30
4 Method	33
4.1 Architecture	33
4.1.1 VAE	33
4.1.2 Discriminator	34
4.2 Training Scheme	34
4.3 Evaluation Metrics	35
5 Results	36
5.1 Human3.6M	36
5.1.1 Evaluation protocol	36
5.1.2 Priliminary Results	36
6 Conclusions	40
6.1 Discussion	40
6.1.1 Future Work	40
6.1.2 Final Words	40

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 Human Pose Estimation (HPE), is a fundamental computer vision task that also forms a basis for advanced tasks such as human action and gesture recognition as well as human motion prediction. 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 the focus of this thesis is on 3D HPE.

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 research explores various ways to solve the task using RGB/Depth image channels, 2D poses, 3D poses, multi-view, and sequential images/videos.

Most of them are supervised learning approaches which require 3D ground truth poses that can only be acquired using physical sensors. Supervised learning methods that learn 3D pose from images, follow a complex cascading approach with 2D poses in some form are an intermediate output. And other learning approaches mostly make use of images from multiple views to estimate the 3D pose.

Assuming 2D poses are obtained from the State-of-The-Art (SOTA) models specialized on 2D HPE models, some works focus on estimating the 3D pose from these 2D poses instead of images. Such networks are called *lifting* networks. These lifting networks can be simple without requiring computationally expensive convolutional layers as it only needs to learn the features of 2D poses that are low dimensional compared to images. However since 2D poses are naturally obtained by projecting 3D poses to a plane, it is an inverse problem. Also, there are multiple feasible 3D poses that when projected result in the same 2D pose. Thus making the task of lifting 2D-to-3D, a *ill-posed inverse* problem due to its inherent ambiguity.

Non-supervised (Weakly/Self/Unsupervised) learning regimes, that are less dependent on 3D pose ground truth, have also gained traction in recent years. Weakly supervised approaches use 3D ground truth indirectly by generating more 2D poses from more views or, for training a discriminator network of a Generative Adversarial Network (GAN). While unsupervised learning (self-supervised) approaches do not use 3D ground truth poses in any shape or form. Many of the deep learning techniques that have already improved the results in other computer vision tasks are yet to be explored in 3D HPE.

1.2 Problem

How can we learn a strong visual representation of the data to tackle the task of 3D HPE? 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 unsupervised learning approaches and 2D-to-3D lifting methods that could help tackle the challenges in scaling 3D HPE to the

real-world.

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, behavior 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 also plays a vital 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 other architectures like autoencoders. These models can use linear, convolutional, and graph layers to learn features. This thesis focuses on investigating the merit in using architectures like Variational Auto-Encoders (VAEs) to solve the 3D HPE within the context of leveraging probabilistic inference models, 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 3D pose 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) and use a different technique to ensure the correctness of pose in 3D. Adversarial training and self-supervision techniques have also given promising results in the last couple of years. The method proposed in the thesis is designed to learn 3D from 2D poses alone in an unsupervised-adversarial learning fashion after the capabilities of the method under supervised settings being verified.

The prime motivation behind the design choices is to address the challenges in scaling up 3D HPE to real-world.

1.6 Stakeholders

Daimler’s ‘Environment Perception for Autonomous Driving’ R&D team in Stuttgart, Germany, 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 understanding human state, motion, and behavior, 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 HPE is a critical element to accomplish this 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 contributions 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 and uses only publicly available, widely used and benchmarked datasets.

1.8 Outline

The theoretical knowledge required to understand the details of the thesis is presented next in chapter 2, Theoretical Background. This entails the explanation of some preliminary concepts 2.1, research area introduction 2.2, where the vast literature related to HPE is summarized, and the highlights of all the works 2.3 that are closely related to the thesis.

The background is followed by chapter 3, Data, providing details of the datasets used along with some visualizations. This chapter also explains the 3D projective geometry concepts required to understand the pre and post-processing steps the data undergoes.

The method, chapter 4, describes the proposed approach in detail. This covers the components of the proposed architecture, the motivation behind the choices, the training and validation procedure and other details that help reproduction.

The results are analysed and discussed in chapter 5 and conclusions are presented in chapter 6.

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.

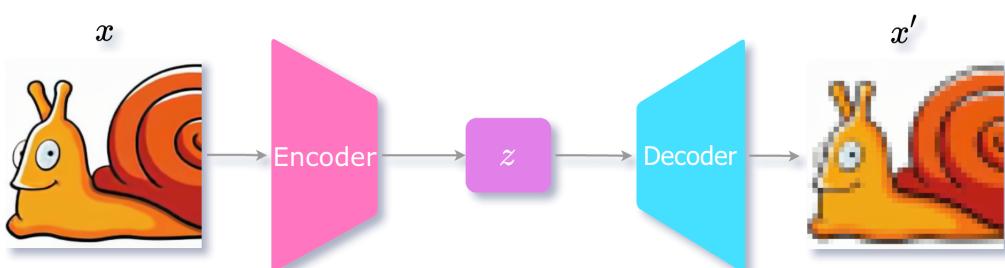


Figure 2.1.1: Illustration of autoencoder architecture. Image source [40]

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 comparatively general properties of the distribution which are 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, 36], learned dimensionality reduction [20]. 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 [41], 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 [40, 47, 49].

2.1.2 Variational Autoencoders

Variational Auto-Encoders (VAEs), unlike standard autoencoders, learn to encode a data sample x as a probabilistic distribution rather than a deterministic value for the latent attribute [25]. 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.

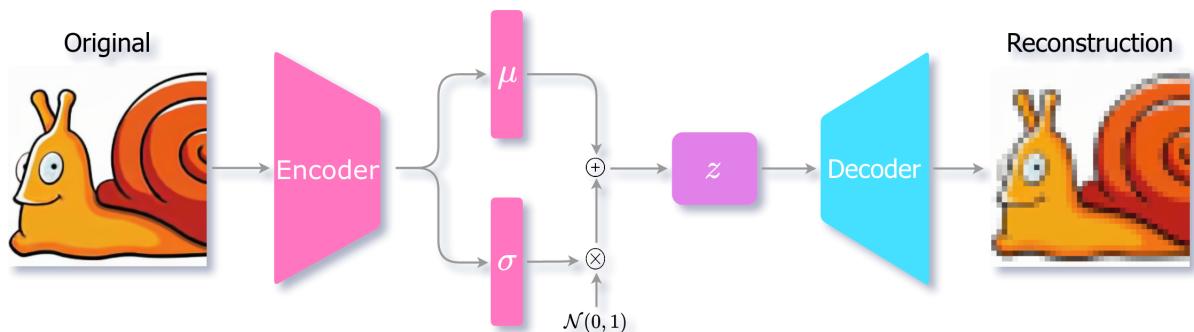


Figure 2.1.2: Illustration of Variational Autoencoder architecture.

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 $p(x)$ Eqn. 2.1 is usually intractable [26].

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)}$$

$$p(x) = \int p(x|z)p(z)dz \quad (2.1)$$

Hence, we try to approximate the posterior distribution by another distribution $q(z|x)$ (the encoder) using variational inference. Variational inference uses optimization to find a distribution that minimizes the Kullback–Leibler Divergence (KLD) to the posterior distribution, $\min D_{\text{KL}}(q(z|x)\|p(z|x))$ while trying to keep the learnt distribution close to the 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:

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

The first term in the above equation makes sure the reconstruction is close to 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}_{VAE} Eqn. 2.3.

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

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 $q_\theta(z|x)$. Where θ is the parameters of the encoder. 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

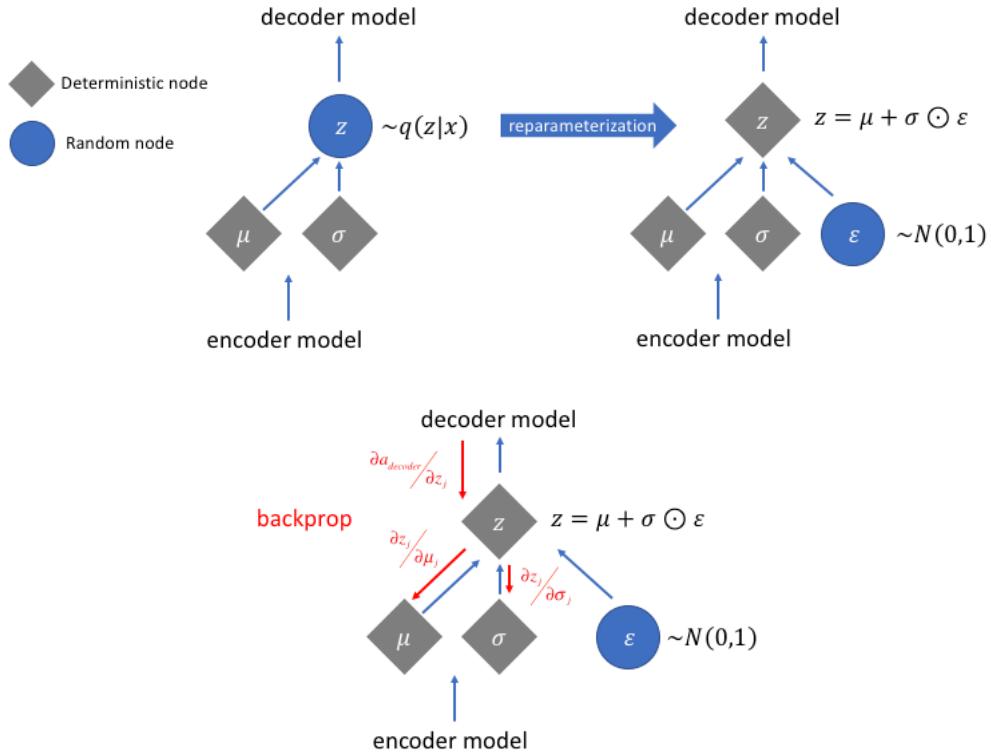


Figure 2.1.3: Comparing the data flow with and without reparametrization trick followed by the backprop calculation. Image source [25]

data from this continuous space, it can generate realistic data by randomly sampling 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.

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.

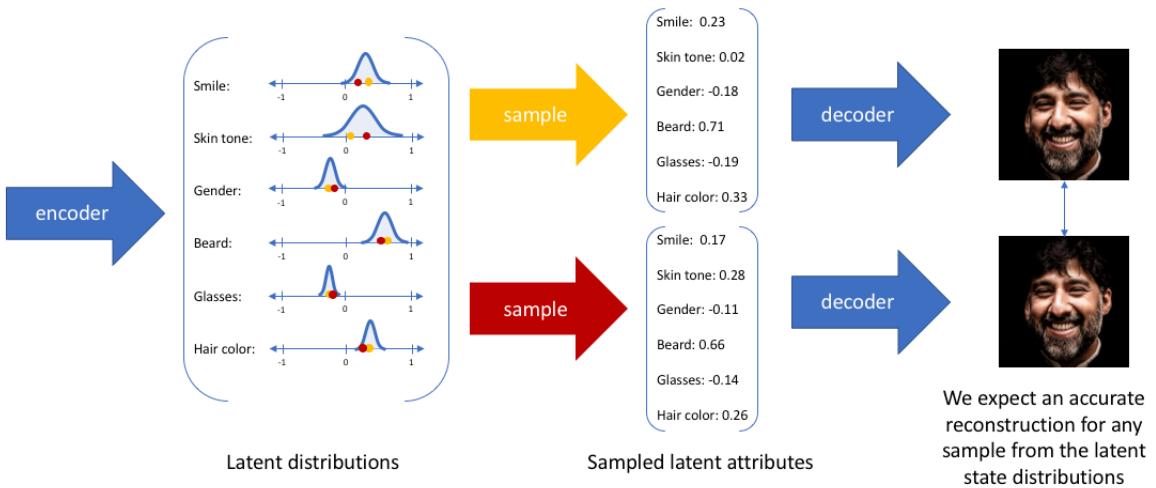


Figure 2.1.4: Probabilistic distribution of latent attributes. Image source [25]

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 Eqn.2.3. So, the loss for the β -VAE is \mathcal{L}_{VAE} Eqn. 2.4. The higher the beta the stronger the constrain on the disentanglement. However, this constraint will negatively affect the representation capability of the VAE.

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

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 as 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

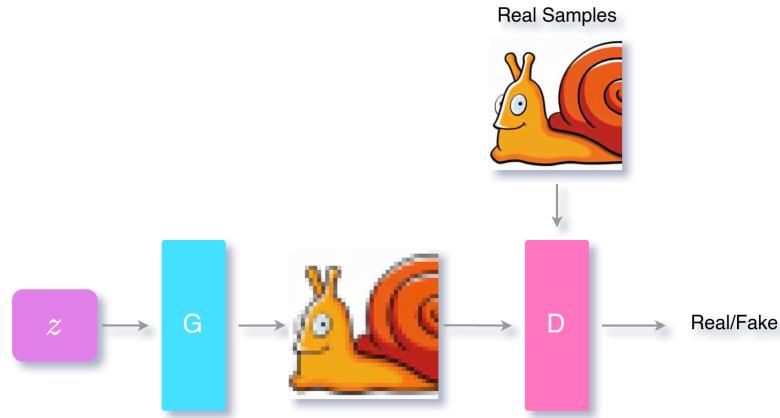


Figure 2.1.5: Illustration of GAN architecture. G is the generator network and D is the discriminator network.

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 $\log(D(x))$ making the loss function as \mathcal{L}_{GAN} Eqn. 2.5.

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

However, when it comes to training a 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 [16]. 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 quickly 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.

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

At global optimality $p_g = p_r$, and for a given generator G , the discriminator at optimality is $D_G^*(x)$ Eqn. 2.6. Hence the virtual cost $C(G)$ [16] when $p_g = p_r$ is:

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

The Jensen–Shannon Divergence (JS-divergence) between two distributions is always non-negative and will be equal to zero only when both the distributions are equal [24]. As we derived in Eqn. 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 [46].

2.1.5 WGAN

Wasserstein Generative Adversarial Networks (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 only 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

JS-divergence used in the standard discriminator based GAN. Since the Wasserstein distance is non-evaluative, a modified version Eqn. 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.

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

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 procedure. 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) [18] 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 [29]. 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.

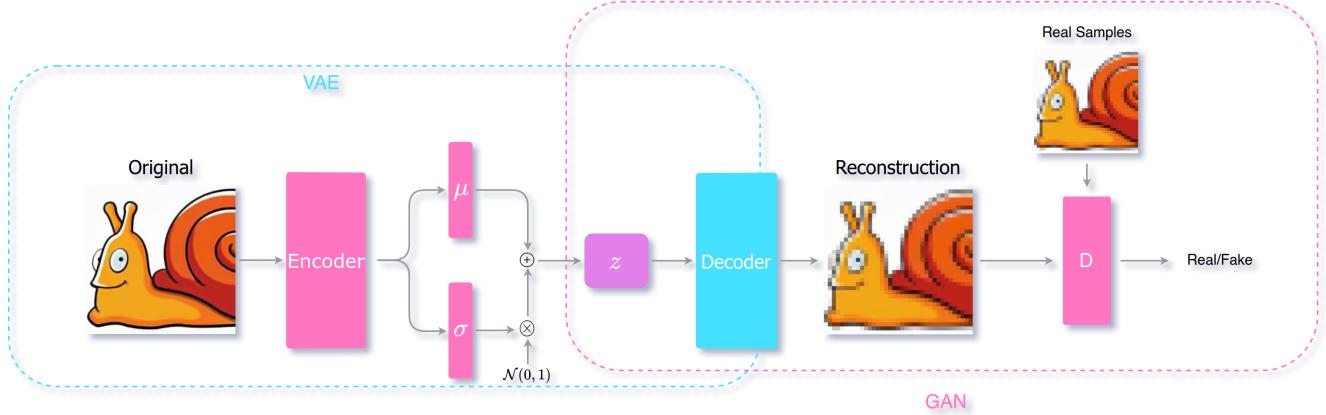


Figure 2.1.6: Illustration of the VAE-GAN architecture

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. 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 Eqn. 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 :

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

Where, \tilde{x} is the generated sample from the VAE's decoder p . $D_l(\tilde{x})$ 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 $\mathcal{L}_{\text{recon}}$:

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

$\mathcal{L}_{\text{recon}}^{D_l}$ 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 Eqn. 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 .

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

Training VAE is hard but training GAN 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.

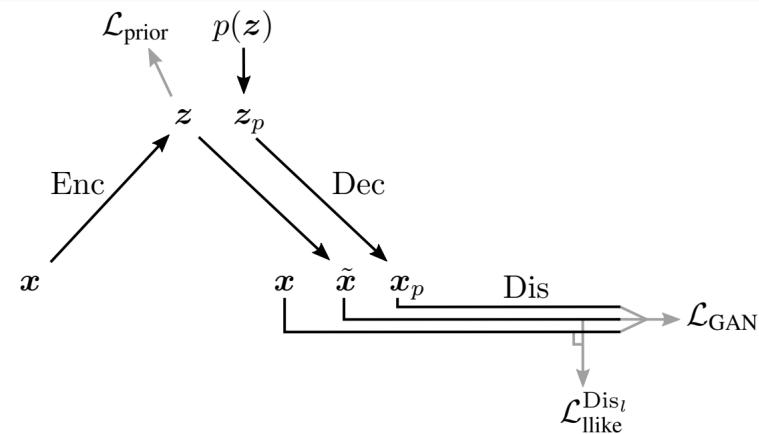


Figure 2.1.7: Illustration of the data flow and the loss of VAE-GAN, where $\mathcal{L}_{\text{recon}}^{D_l} = \mathcal{L}_{\text{llike}}^{Dis_l}$
Image source [29]

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 between the quality of the reconstruction 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}_{GAN} as in Eqn. 2.12.

$$\theta_p \leftarrow t - \nabla_{\theta_p} (\gamma L_{\text{uike}}^D - \mathcal{L}_{\text{GAN}}) \quad (2.12)$$

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

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

2.2 Research Area Introduction

In this section, the details of various approaches and the State-of-The-Art (SOTA) in 3D Human Pose Estimation (HPE) are presented along with some works in the sibling tasks of hand pose estimation. This section aims to only give an overview of how the problem of 3D HPE is tackled in the literature, approaches that are directly related to the method proposed in this thesis are presented later in section 2.3, Related Works. The different categories of the approaches mentioned below are not exclusive but are the main aspects of the described approaches.

2.2.1 Cascading Approach

Numerous works try to estimate 3D human poses from 2D RGB images or 2D joint confidence heatmaps [7, 9, 34, 37, 44]. Most of these methods follow a cascading approach, where an explicit intermediate representation of 2D heatmaps or 2D poses are predicted. This typically involves training a convolutional neural network to find humans in the image or just extract features that correspond to the joints. These features are forwarded to another neural network that learns to estimate the corresponding 3D pose, usually by predicting the depth offsets of the 2D joint or sometimes by predicting the shape and the view parameters of the 3D pose.

For example, [34] proposes a general framework with 3 networks as depicted in Fig 2.2.1. 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 the possibility to divide the task of RGB to 3D into smaller, well-studied sub-tasks. This enables explicit supervision of known intermediate states that could be of interest for understanding the representations learned by the network or to use the intermediate output for other auxiliary tasks. Moreover, in this case it 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. Most importantly certain modules can be replaced or improvised without affecting or having to re-train the entire pipeline.

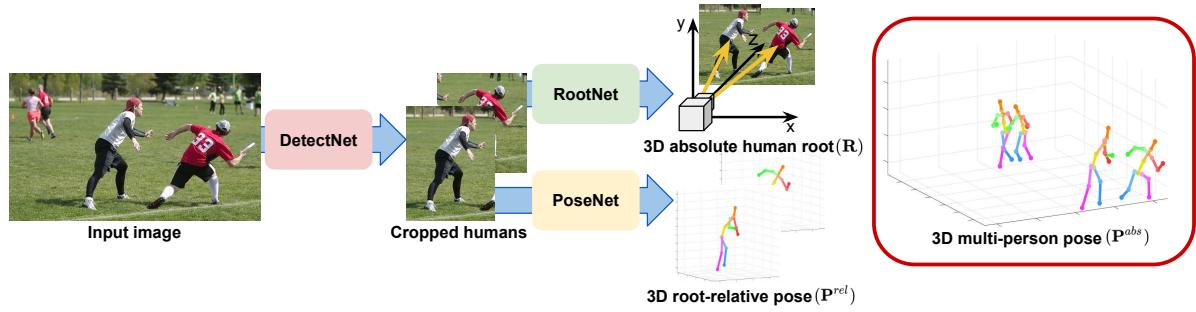


Figure 2.2.1: Pipeline of an absolute 3D human pose estimation network following a cascading approach [34], showing different neural networks such as DetectNet, RootNet and PoseNet that trained on specific subtasks to estimate absolute 3D human pose.

2.2.2 Pose Lifting

In contrast to the estimating pose from an image, Pose Lifting works such as [7, 8, 28, 35, 43], focus on estimating 3D poses from 2D poses alone while assuming 2D poses from the SOTA methods in 2D HPE. This category follows the ideology of cascading based approaches but restricts the study and discussion purely to lifting 2D to 3D pose. These methods include simple linear models as first described in [33] with a series of fully connected linear, batch normalization, dropout layers with residual connections as illustrated in Fig 2.2.2 to regress 3D pose effectively. These simple networks have enough capability to capture the features of the data as the input and output data are much smaller in dimensions as compared to RGB images.

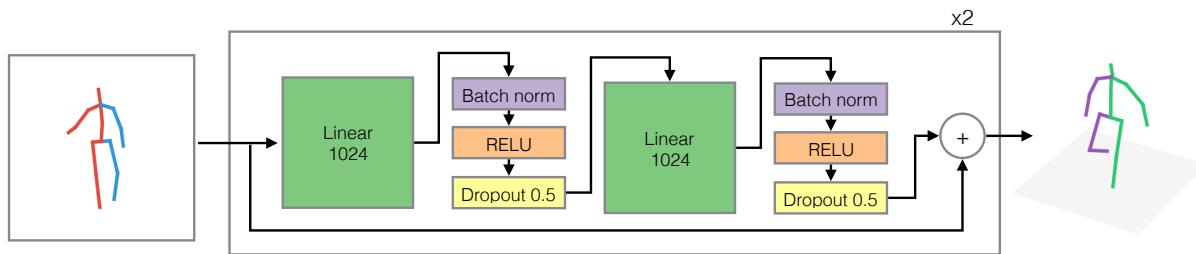


Figure 2.2.2: Architecture of the first deep learning based lifting approach proposed by Martinez *et al.* [33]

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 projections (2D keypoints). This approach has been widely used in facial keypoint detection and [27] introduces deep learning variant for the same. Instead of predicting the 3D coordinates of each keypoint/joint of

the 3D pose, [27, 35, 44, 45] predicts the 3D shape and camera pose from 2D pose using this method. The advantage of such approach is to obtain disentangled representation of view and 3D pose that is invariant of the view or camera position.

The Lifting approaches facilitate 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 computationally inexpensive for both training and inference. Moreover, the 2D and 3D pose data usually can be entirely loaded onto the GPU further accelerating the training procedure. Thus addressing one of the major problems that effects the scalability of 3D HPE models as well as enabling the development of better modular systems by combining the best of Lifting networks with the best of 2D HPE.

2.2.3 Multiple Hypothesis Estimation

As stated in 1.1, 2D-to-3D pose lifting is an ill-posed-inversed problem due to inherent depth ambiguity as there are multiple plausible 3D poses that gives the same 2D projection, this shall be further elaborated later in chapter 3, Data. To address this problem, Jahangiri *et al.* [23] propose a solution a 3D Gaussian Mixture Model (GMM) to learn uniformly sampled 3D poses and conditionally sample to retrieve 3D poses that have the reprojection error within the given limits. Thus estimating multiple 3D poses conditioning on a single input 2D pose, similar to a dictionary based learning approach.

More deep learning based approaches such as [30, 31, 37] were later introduced. Out of them, Chen Li *et al.* [30] proposes a variational inference model replacing the GMM with a Mixture Density Network (MDN) which was first introduced by Ye *et al.* [48] to handle occlusions in hand pose estimation. Thus addressing two of the major problems of 3D HPE i.e missing joints from occlusions and variational inference.

Another interesting approach from Sharma *et al.* [37], presents a conditional Variational Auto-Encoder which takes a random sample from a normal distinguish as an input and 2D pose as a condition and outputs different 3D pose for different random samples given the same 2D pose. This approach also handles missing joints and provides variational inference. A evaluation technique is also proposed to rank each of the multiple hypotheses by scoring the poses based on joint-ordinal depth

relations learnt from the images or by oracle score that access 3D ground truth to compute the closest match as illustrated in Fig 2.2.3.

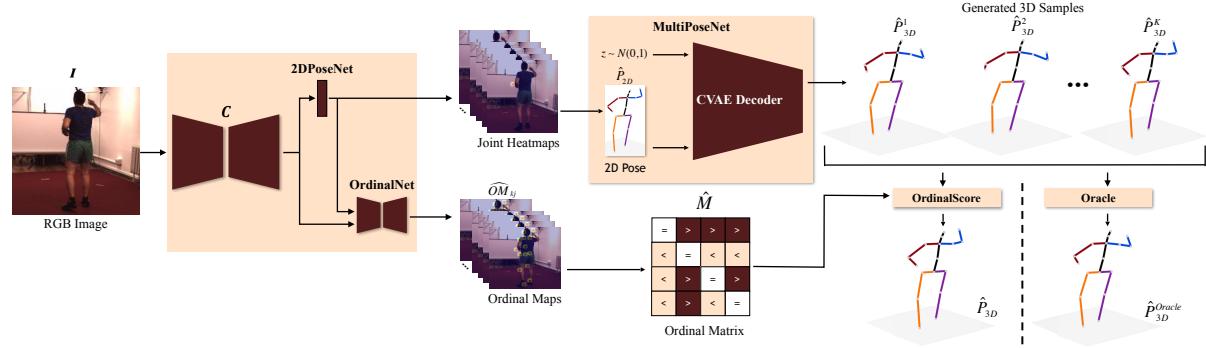


Figure 2.2.3: Illustration of modular framework from [37] that uses a Conditional VAE to estimate multiple hypothesis for a given 2D pose as condition. The ordinal scoring learnt from joint-depth ordinal relations and the score obtained from the oracle that learns scores higher to poses closest to the ground truth pose are used rank to each of the predicted 3D pose.

Recent work from Chen Li *et al.* [31] which is an improvised version of their earlier work [30] proposes a weakly supervised approach that is much more similar to that of this thesis. The details are discussed later in related works section 2.3. However, all of the mentioned approaches still require 3D poses in one way or other for training and hence do not address the most important bottleneck of obtaining 3D ground truth.

2.2.4 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 3D 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 [22], Human Eva [21] and, multi-person datasets such as CMU Panoptic [19] provide 3D pose ground truth, they are obtained using Motion Capture (MoCap) systems Fig[2.2.4] which are only limited to indoors or cannot be directly adapted to outdoor environments where the majority of the use cases exist. 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. The datasets from simulations come with

the difficulty of domain adaptation to be transferable to the real world.



Figure 2.2.4: Image from Human3.6 Dataset [22] of subject wearing MoCap markers

To overcome this bottleneck, [28] proposes unsupervised training of a generative 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 as illustrated in Fig 2.2.5. Thus removing the need for any explicit 3D annotations besides 2D pose that are either manually labeled or obtained using 2D HPE models. RepNet [43] 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. However, due to the inherent ambiguity in lifting 2D pose to 3D and as the images are not captured with orthogonal cameras, reprojection of 3D

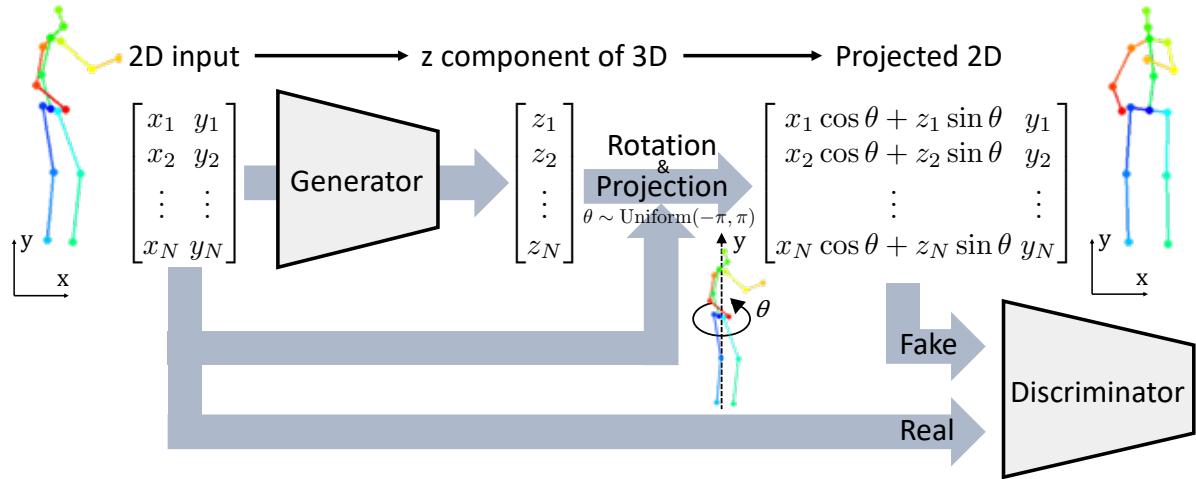


Figure 2.2.5: Architecture of a simple unsupervised adversarial learning architecture proposed by [28]. The network takes 2D coordinates (x, y) of each joint in the pose and predicts the corresponding z -component. The 3D pose achieved by combining the input x, y with the predicted z , is the randomly rotated along the y -axis to retrieve a novel view of the 3D pose. This pose is the projected to xy -plane, which would be much different from the input 2D pose. This projected 2D pose should be similar to the real 2D poses in the dataset if the predicted 3D represents true human pose. Their similarity to real 2D poses is reinforced by a discriminator. The whole training procedure is carried out without the need for 3D pose in any shape or form.

pose is not necessarily consistent with the ground truth as the camera intrinsic and extrinsic parameters are not taken into account. Hence it is challenging to match the performance of models trained on 3D ground truth.

2.2.5 Multimodal Representation Learning

Another interesting approach is training VAEs using multiple modalities like images, poses, depth maps [17, 38, 39, 42]. Multimodal Variational Auto-Encoder (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[2.2.6]. 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.

Mixture-of-Experts Multimodal Variational Auto-Encoder (MMVAE) proposed by

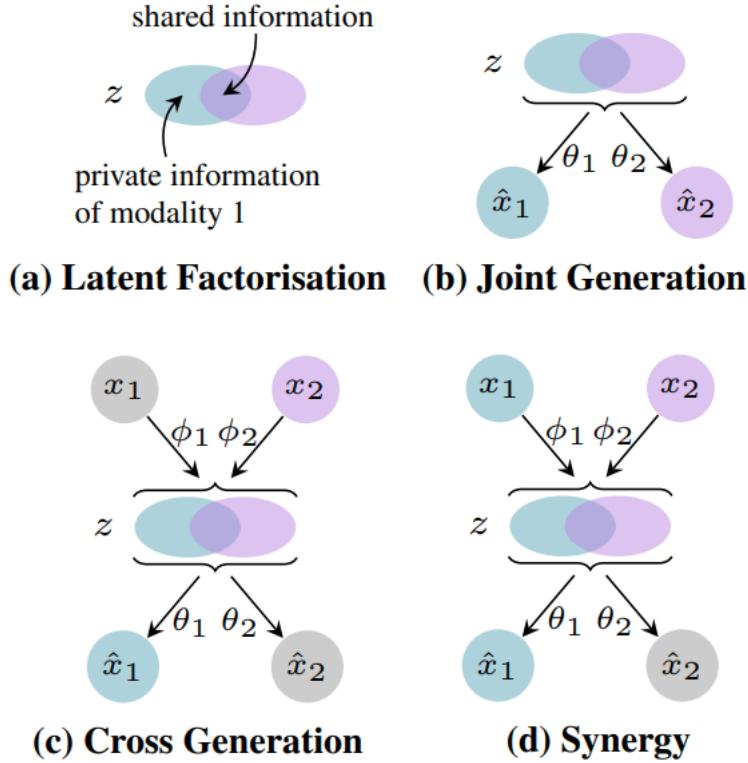


Figure 2.2.6: Criteria for Multimodal Generation [38]

[38] 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 [39]. 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

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, 28, 35] are the main approaches that do not use 3D supervision in any way. While [31, 43] 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 or most likely published after finishing this thesis report.

[8, 13, 28] 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 earlier 2.2.4, 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 have 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.* [43] also mentioned 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 parameter 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 of **depth ambiguity** persists.

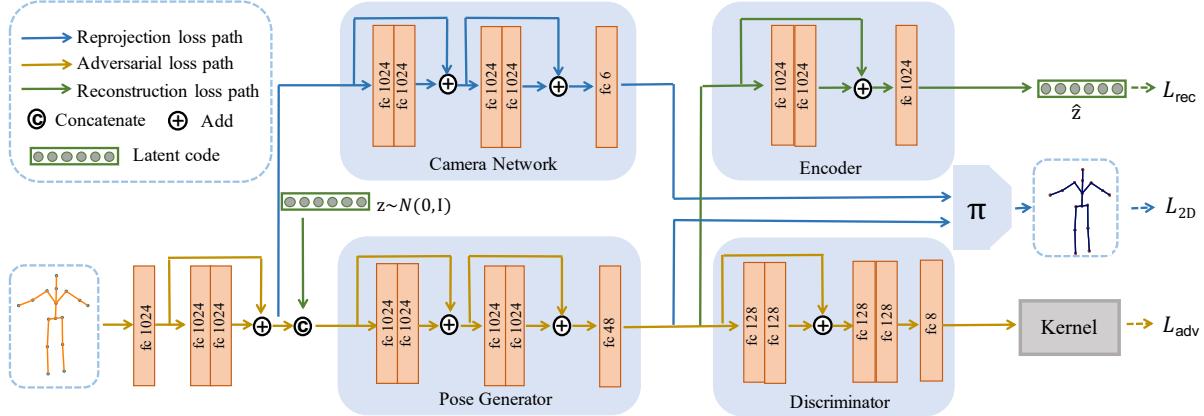


Figure 2.3.1: Illustration of the weakly supervised multiple hypothesis generation architecture proposed by [31]

Chen Li *et al.* [31], building upon their previous work [30] that is explained earlier in 2.2.3, proposes a weakly supervised variational inference model as the proposed method in this thesis was developed. The new approach [31] was inspired by the architecture of Wandt *et al.* [43] that does not require direct 3D supervision. This model first encodes the input 2D pose to a latent representation. This representation is then concatenated with a *latent code* to produce a 3D pose as well as used by the camera network to predict the camera pose. Instead of directly supervising the predicted 3D pose with the ground truth, it is projected back to 2D using the predicted camera view which is self supervised as it should be the same as the input 2D pose.

This self supervision trains the network to only output 3D poses that are close to input 2D in a particular view but unconstrained from other view. To ensure, the predicted 3D is feasible and close to ground truth a WGAN is used as illustrated in Fig 2.3.1. The predicted 3D pose is passed to a critic network that scores the poses on how realistic they are. This encourages the pose network to generate poses that are realistic, in other words indistinguishable from the 3D ground truth poses. This results in a realistic 3D pose that is close to 2D input pose which is most likely close to its corresponding 3D ground truth. In addition to this an encoder is trained to reconstruct the latent code to

ensure diversity and prevent mode collapse of the WGAN.

This variational inference of 3D pose addresses both the problems of depth ambiguity and missing joints. However, this weakly supervised still requires 3D ground truth poses to train the discriminator and does not address the problem completely. The method proposed in this thesis is similar to this approach but does not require 3D in any shape or form, and uses much smaller architecture without the camera and latent vector encoder networks while addressing all the major problems.

Canonical 3D Pose Networks for Non-Rigid Structure From Motion (C3DPO) [35], which is also discussed in 2.2.4 handles missing joint and is fully unsupervised but does not have the capability of predicting multiple hypotheses for a given 2D pose. This approach that uses NRSfM in an unsupervised way, does not yield good results compared to other approaches. In this thesis, we present a method that has the merits of the above approaches and address all the aforementioned problems.

Chapter 3

Data

This chapter discusses the datasets used in the thesis, as well as the pre and post-processing steps that are performed in order to facilitate the training and validation of the proposed method. In addition to this, a couple of concepts that motivate the choices or that are directly used in the process of the data are explained with the help of illustrations.

The main dataset used in the thesis is *Human3.6M*: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments [22]. 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 [32] would be used.

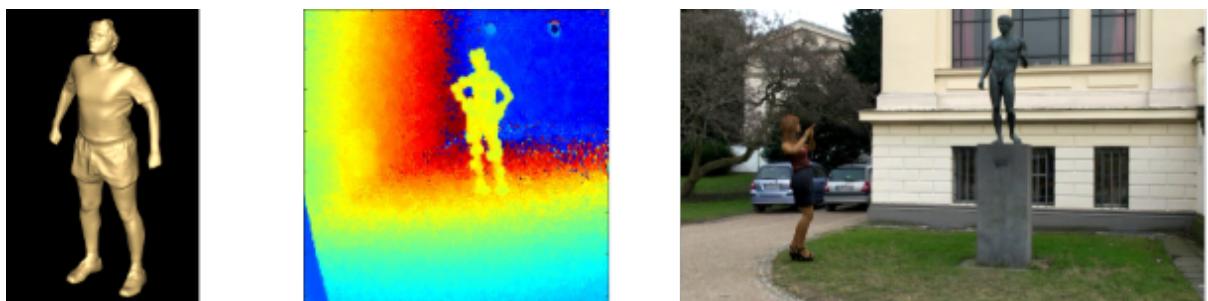


Figure 3.0.1: Additional modalities such as (from left to right) Full body model, depth from time of flight and mixed reality data are available in Human3.6M dataset [22]. These datasets can be used for human body estimation, depth map estimation or inferring absolute 2D or 3D pose directly from full image i.e without cropping the region with a single person.

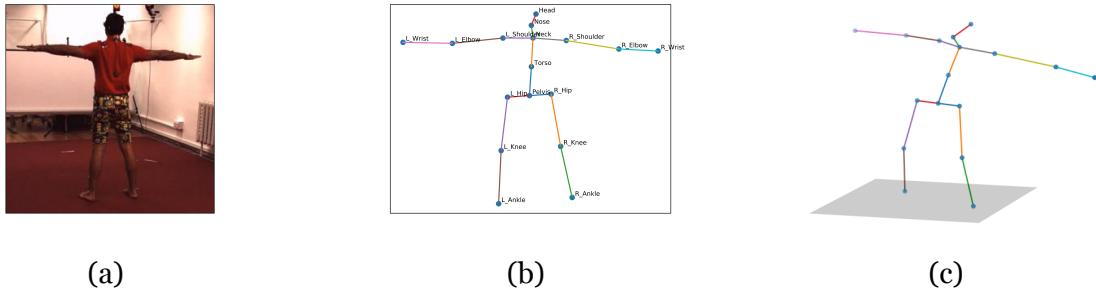


Figure 3.1.1: (a) RGB image from one of the 4 cameras. (b) Corresponding 2D pose obtained from the 3D pose to the right with joint labels. (c) Corresponding 3D pose, rotated for a better view of the spatial distribution of the joints

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 marker-based Motion Capture (MoCap) system. The dataset constitutes 15 diverse motion and actions in various everyday scenarios namely, Directions, Discussion, Eating, Greeting, Phoning, Photo, Posing, Purchases, Sitting, Sitting Down, Smoking, Waiting, WalkDog, Walking, and WalkTogether. These actions are performed by 11 professional actors wearing a variety of realistic clothing. The datasets provides synchronized 2D and 3D data including full-body scans as shown in Fig 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.

The data of interest is mainly the 2D pose for training, 3D poses for evaluation and images for qualitative analysis as it is sometimes challenging even for the human eye to estimate 3D pose just from the 2D skeleton. Both the poses are composed of 17 joints or keypoint annotations namely, Pelvis (also referred to as Root), Torso, Neck, Nose, Head, right and left - Hip, Knee, Ankle, Shoulder, Elbow, Wrist. An image sample from the dataset with its corresponding 2D and 3D pose is illustrated in the Fig 3.1.1.

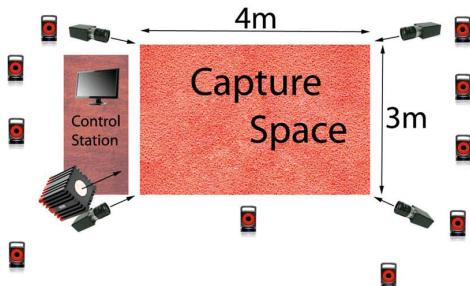
As illustrated in Fig 3.1.2, the data is collected from a indoor environment with multiple cameras and MoCap sensors. The obtained coordinates of the 3D pose from MoCap data are global and is relative to a fixed origin the recording room. In addition to the global 3D pose, camera relative, i.e local 3D pose is also obtained using the camera's location with respect to the global origin. The global poses are useful for methods that try to exploit the multiview information and local poses that are relative to the camera

can be used for absolute pose estimation. The 2D pose from each view is obtained by projection the 3D pose using the parameters of the respective camera. This 2D pose is relative to that particular camera and is with respect to the full-scale image that captures the entire scene in the camera's field of view. Along with the RGB images, global and local 3D pose and 2D pose, other metadata is available. This metadata includes intrinsic and extrinsic parameters of each camera, an identifier for each of the samples of a particular action sequence, bounding box coordinates of the human in the full-scale image, subject, action, subaction and camera ids for each sample.

Type of action	Scenarios	Train	Validation	Test	MoCap System	DV System		
Upper body movement	Directions Discussion	83,856 154,392	50,808 68,640	114,080 140,764	No x Sensor Resolution Freq. Sync	10 x Vicon T40 4 Megapixels 200Hz hardware	No x Sensor Resolution Freq. Sync	4 x Basler piA1000 1000x1000 50Hz hardware
Full body upright variations	Greeting Posing Purchases Taking Photo Waiting	69,984 70,948 49,096 67,152 98,232	33,096 25,800 33,268 38,216 54,928	84,980 85,912 48,496 89,608 123,432	TOF System	Body Scanner		
Walking variations	Walking Walking Dog Walking Pair	114,468 77,068 76,620	47,540 30,648 36,876	93,320 59,032 52,724	No x Sensor Resolution Freq. Sync	1 x Mesa SR4000 176x144 25Hz software	Sensor No. Lasers Point Density Tolerance	Vitus Smart LC3 3 7dots/cm3 < 1mm
Variations while seated on a chair	Eating Phone Talk Sitting Smoking	109,360 132,612 110,228 138,028	39,372 39,308 46,520 50,776	97,192 92,036 89,616 85,520				
Sitting on the floor	Sitting Down	112,172	50,384	105,396				
Various Movements	Miscellaneous	-	-	105,576				
Total		1,464,216	646,180	1,467,684				

(a) The number of 3D human poses in Human3.6M in training, validation and testing aggregated over each scenario. We used 5 subjects for training (2 female and 3 male), 2 for validation (1 female and 1 male) and 4 subjects for testing (2 female and 2 male). The number of video frames is the same as the number of poses (4 cameras capturing at 50Hz). The number of TOF frames can be obtained by dividing the table entries by 8 (1 sensor capturing at 25Hz).

(b) Technical summary of our different sensors.



(c) Floor plan showing the capture region and the placement of the video, MoCap and TOF cameras.

Figure 3.1.2: Human3.6M data collection details. Image source [22]

3.2 Projective Geometry

The projection of the poses in the thesis is using a simple pinhole camera model as illustrated in fig.3.2.1. 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.

3.3 Depth Ambiguity and Camera Modeling

One of the main problems discussed in ?? is depth ambiguity. The fig ?? 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

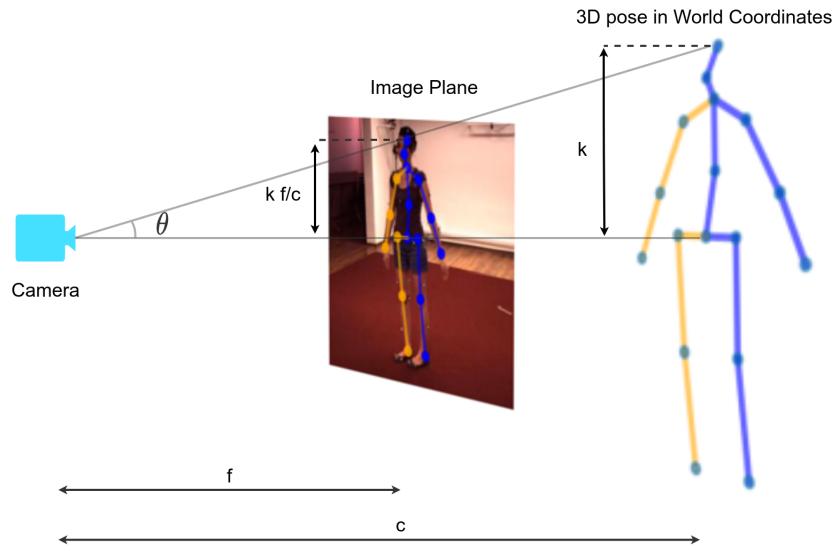


Figure 3.2.1: Pinhole Camera Model. Image Source [2]

as illustrated in fig. 3.4.1. But in the proposed method we only translated and rotate but do not scale the pose.

3.4 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.

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.

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 the image plane is fixed. As

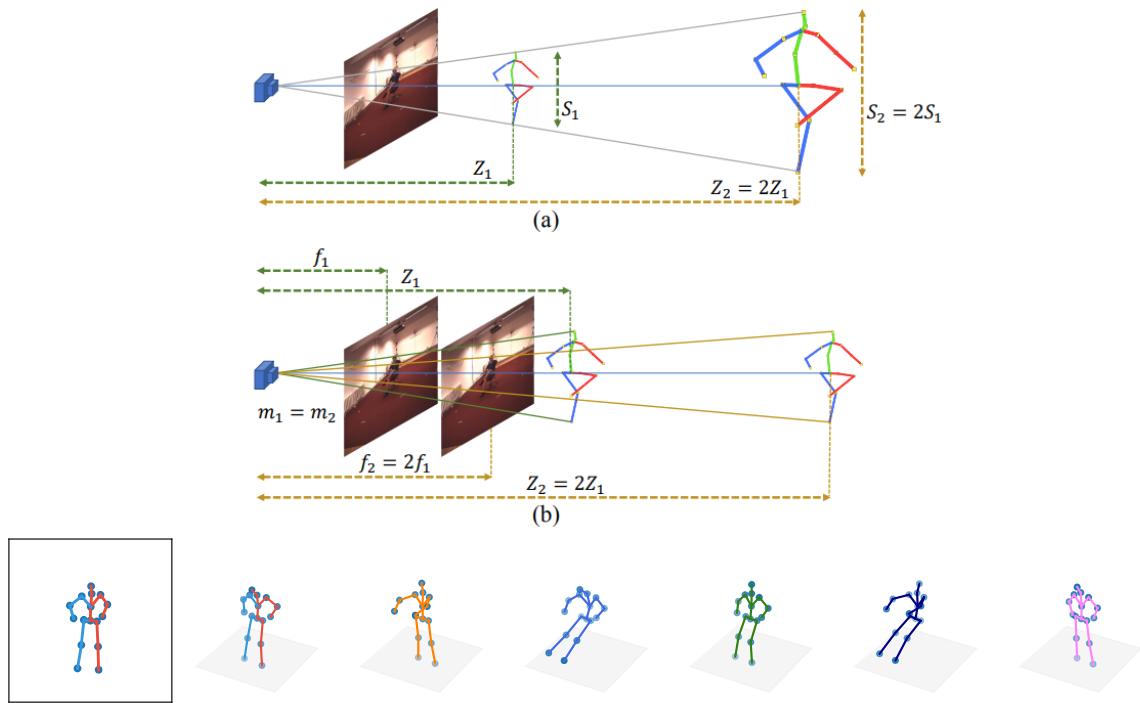
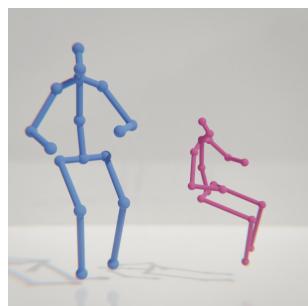


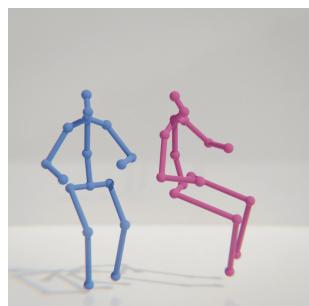
Figure 3.3.1: shows 2 of the infinite possible poses that result in the same 2D reprojection. Where (b) shows the same phenomenon for different focal lengths Image Source [7]

a unit camera or camera with unit focal length is assumed the projection of unit 3D poses onto a plane at a fixed distance of $1/c$ units. The root of the 2D pose is however removed to remain consistent with the 3D pose.

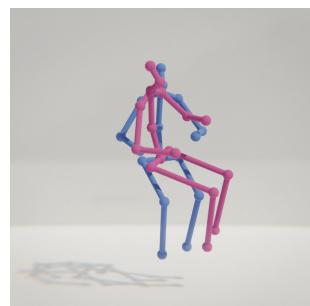
The estimated poses from the networks that are trained on downsampled poses are upscaled to the original size using for valuation. This postprocessing step is required for getting the distance between prediction and ground truth keypoints in true units of millimeters.



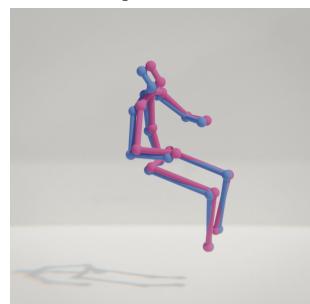
(a) $y = x$



(b) $y = 3\sin x$



(c) $y = 5/x$



(d) $y = 5/x$

Figure 3.4.1: Three simple graphs

Chapter 4

Method

The method used in the thesis is the Variational Auto-Encoder (VAE)-Generative Adversarial Network (GAN) hybrid discussed in 2.1.6. Except that the Similarity function used is Least Absolute Deviations (L1) loss and not the GAN similarity metric 2.9. The details are further described in the section.

4.1 Architecture

In this section, the components of the proposed network and the design motivations are described. The VAE-GAN hybrid architecture is illustrated in fig. 4.1.1

4.1.1 VAE

Similar to the VAE-GAN described in 2.1.6, the proposed architecture consists of an encoder, decoder and a discriminator. The encoder takes the 2D pose as input and encodes it in a latent space. The decoder samples this encoding using the reparametrization trick and outputs a 3D pose. Following the best practices from [4], the decoder which is also the generator is made to output 3D poses with values from [-1,1] and Tanh activation is used at the output of the decoder. As described in 3, the 3D pose is modeled to be of unit length from root to head. However, the lower half of the body could be higher than that. To handle this, the output is multiplied by 2 to cover the true dimensions of the lower half.

The encoder consists of an upsampling layer that makes the $2 \cdot j$ dimensional input to match the number of hidden neurons of the encoding module. Where j is the number

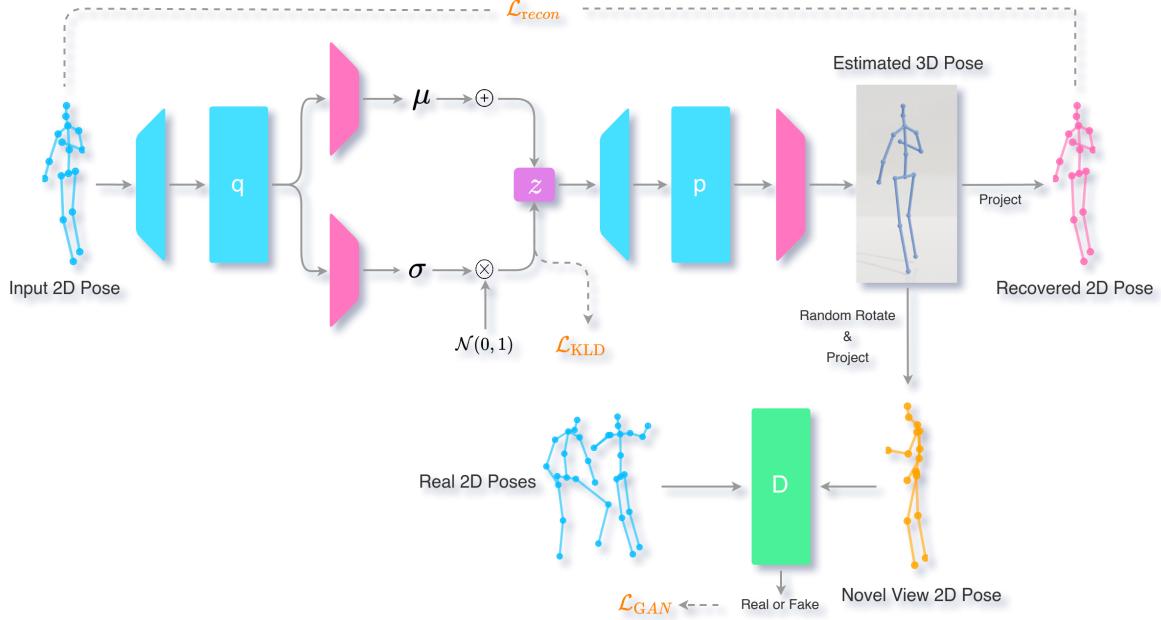


Figure 4.1.1: Illustration of the architecutre and loss flow in the proposed approach.

of joints, here 16. The encoding module is made of n residual block composed of 2 Fully Connected (FC) layers following the related works, to allow comparison. Where n is usually 1 or 2. The enoding block is followed by 2 FC layers that downsample the hidden representation to match the latent space dimension. The output of the two downsampling layers represents the mean and variance of the embedding in the latent space. All the layers use Rectified Linear Unit (ReLU) activation unless specified otherwise. The decoder mimics the encoder with an upsampling layer, a decoding module, and a downsampling layer to output 3D pose of dimension $3 \cdot j$.

4.1.2 Discriminator

The discriminator following the related works also mimics the generator and takes 2D poses as input and predicts binary labels of real or false. The sigmoid activation function is used at the output layer of the discriminator.

4.2 Training Scheme

The training scheme is similar to the standard VAE and VAE-GAN introduced in 2.1. The discriminator is trained first for a few steps before training the VAE. First, the VAE takes the 2D pose as input and predicts the 3D pose. This predicted 3D pose is

first reprojected to 2D to train the VAE and then randomly rotated and reprojected to a novel 2D view. This novel 2D is used as fake samples to train the discriminator. As VAE learns to improve the reprojected 2D pose, it will eventually be very close to the real samples. Using these as the fake examples will not be very useful as the decoder does the same job. Training the discriminator on a randomly rotated and reprojected 2D novel view would encourage the decoder to generate a pose that not only agrees with the input 2D pose but ensures that the other view of the 3D is also indistinguishable from the variations found in the data by fooling the discriminator. This two-view supervision leads to better 3D pose generations.

4.3 Evaluation Metrics

3D Human Pose Estimation (HPE) and Human3.6M in particular is mainly evaluated by Mean Per Joint Position Estimate (MPJPE) metric. MPJPE as it 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 protocols. Protocol-1 is using all 4 camera views in subjects S_1 , S_5 , S_6 , S_7 , and S_8 for training and the same 4 camera views in subjects S_9 and S_{11} for validation/testing. Protocol-2 is the same as 1 except that the predictions are post-processed via a rigid transformation before comparing to the ground-truth.

5.1.2 Preliminary Results

The results presented are after training the model for ~ 1400 epochs on approximately 300,000 2D poses with a batch size of 256 on a Titan X while flipping them with a probability of 0.5. The architecture is as described earlier with 1024 hidden units per linear layer and 512 latent dimensions. Both the Variational Auto-Encoder (VAE) and the discriminator are trained using Adam optimizer with default hyperparameters and with a learning rate of 2e-4.

The different poses during the training are presented in fig. 5.1.1. The poses in pink are the ground truth while the ones in blue are the predictions. The poses refer to ground truth 2D, reconstruction 2D, reconstruction novel view, reconstruction 3D, and combined reconstruction and ground truth 3D after alignment. The same order is

followed for the other visualization.

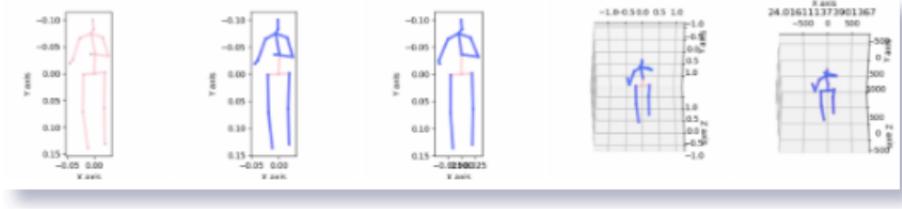


Figure 5.1.1: Visualization of the inputs and outputs of the model

One of the challenging parts is finding the optimal weights for each of the terms in the triplet loss. The β value for the Beta Variational Auto-Encoder (β -VAE) is cycled from 0 to 0.001 for every 40 epochs while keeping it constant at 0.001 for 10 epochs with a 10 epoch warmup at the beginning of the training. The weight of the critic with scaling the magnitudes is 1e-4. While the weight for the reconstruction loss is kept at 1.

The discriminator is trained for 3 iterations for one iteration of the VAE. The training is noise and the networks need to train for several hundreds of epochs before they tend to converge. The experiment which the results are from is trained over 1400 epochs but very slow improvements in the results are observed. At 1400 epochs the Mean Per Joint Position Estimate (MPJPE) is ~ 68 mm. This is far less than the current state of the art [8] with ~ 40 mm. The results are believed to improve using Wasserstein Generative Adversarial Networks (WGANs) as they offer better convergence.

Despite the MPJPE being ~ 68 , the decoder generates many accurate 3D poses that are almost indistinguishable for the human eye. The high number of outliers and lower performance on hard poses worsens the overall performance of the model. The graph on the left depicted in fig 5.1.3 shows the slow and gradual decrease in MPJPE. And the graph on the right shows the histogram of the Per Joint Position Estimate (PJPE) of each sample over time. This shows the model is unable to certain poses while it improves gradually on the rest.

Some of such outliers are presented in fig 5.1.2, the predictions in (a) are the ones the model is unable to learn. While (b) is the evidence of the shortcoming of the current processing technique. Rectifying that would improve the evaluation metric of the model quite significantly.

The visualization of 2D pose embedding in latent space after dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) is shown in fig. 5.1.4.

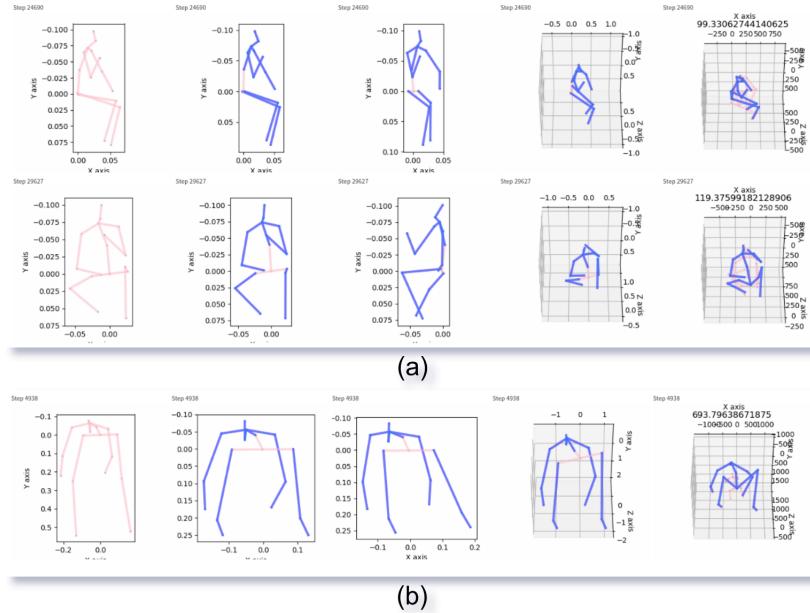


Figure 5.1.2: (a) Prediction on hard poses with high ambiguity. (b) Poses that can be improved with changes to data processing.

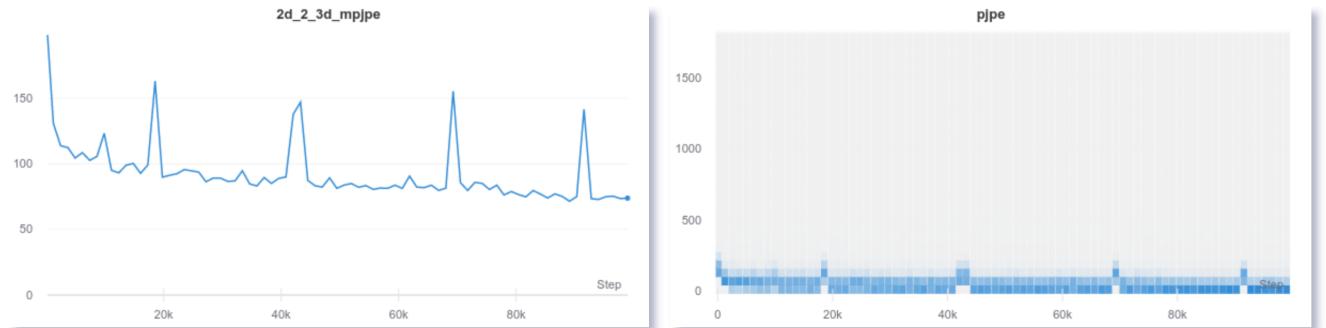


Figure 5.1.3: MPJPE and PJPE trends during the training respectively

Each action is given a unique color. Though we see small clusters of blues, browns, pinks the overall space looks very mixed up. This is expected as many of the instances in different actions overlap. For example, the action standing up and sitting down have instances while both are standing or sitting, etc.

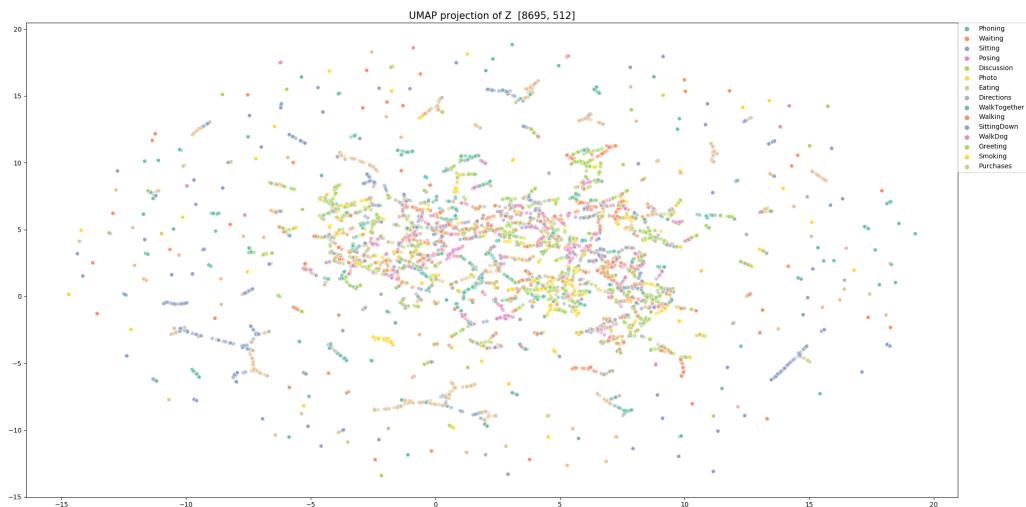


Figure 5.1.4: UMAP Visualization of samples in latent space. The actions do not always directly related to the pose due to overlaps from one action to another. Better viewed in color and 200% zoom

Chapter 6

Conclusions

YET TO BE WRITTEN

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.

6.1 Discussion

6.1.1 Future Work

6.1.2 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] 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].
- [17] 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>.
- [18] Gulrajani, Ishaan, Ahmed, Faruk, Arjovsky, Martin, Dumoulin, Vincent, and Courville, Aaron. *Improved Training of Wasserstein GANs*. 2017. arXiv: 1704.00028 [cs.LG].

BIBLIOGRAPHY

- [19] Hanbyul Joo, Hao Liu. “Panoptic Studio: A Massively Multiview System for Social Motion Capture”. In: (2015).
- [20] 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>.
- [21] *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).
- [22] 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.
- [23] Jahangiri, Ehsan and Yuille, Alan L. *Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections*. 2017. arXiv: 1702.02258 [cs.CV].
- [24] *Jensen–Shannon divergence*. Sept. 2020. URL: https://en.wikipedia.org/wiki/Jensen%20Shannon_divergence.
- [25] Jordan, Jeremy. *Variational autoencoders*. July 2018. URL: <https://www.jeremyjordan.me/variational-autoencoders/>.
- [26] Kingma, Diederik P and Welling, Max. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML].
- [27] 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>.
- [28] 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>.

- [29] 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].
- [30] Li, Chen and Lee, Gim Hee. *Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network*. 2019. arXiv: 1904 . 05547 [cs.CV].
- [31] Li, Chen and Lee, Gim Hee. *Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses*. 2020. arXiv: 2008.05770 [cs.CV].
- [32] 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.
- [33] 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>.
- [34] 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>.
- [35] 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.
- [36] 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.
- [37] 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].

- [38] 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>.
- [39] 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_cvpr_2018/html/Spurr_Cross-Modal_Deep_Variational_CVPR_2018_paper.html.
- [40] Ulyanov, Dmitry, Vedaldi, Andrea, and Lempitsky, Victor. “Deep Image Prior”. In: *arXiv:1711.10925* (2017).
- [41] 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.
- [42] 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>.
- [43] 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>.
- [44] 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>.

BIBLIOGRAPHY

- [45] 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>.
- [46] Weng, Lilian. *From GAN to WGAN*. 2019. arXiv: 1904.08994 [cs.LG].
- [47] 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.
- [48] Ye, Qi and Kim, Tae-Kyun. *Occlusion-aware Hand Pose Estimation Using Hierarchical Mixture Density Network*. 2018. arXiv: 1711.10872 [cs.CV].
- [49] Zhang, Richard, Isola, Phillip, and Efros, Alexei A. “Colorful image colorization”. In: *European conference on computer vision*. Springer. 2016, pp. 649–666.

Appendix - Contents

A First Appendix	42
B Second Appendix	43
References	41

Appendix A

First Appendix

**** YET TO BE WRITTEN ****

This is only slightly related to the rest of the report

Appendix B

Second Appendix

this is the information