UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Controllable synthetic image datasets generation for advancements in human pose and shape estimation

Author:
Ostap VINIAVSKYI

Supervisor:
Orest KUPYN

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Ostap VINIAVSKYI, declare that this thesis titled, "Controllable synthetic image datasets generation for advancements in human pose and shape estimation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"There is no globally-optimal life. There is no sequence of choices in life that will produce the 'perfect life' or 'perfect career'. This is hard to accept, but once you accept it, it's very freeing.
So my advice is to choose the option that is the most ***fun****."*

Michael J. Black

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Controllable synthetic image datasets generation for advancements in human pose and shape estimation

by Ostap VINIAVSKYI

Abstract

Human-centric applications are ubiquitous in the modern world. Myriads of educational, entertainment, e-commerce, and other applications require understanding the measurements of the human body in the image. The deep-learning methods for solving human-centric problems usually rely on supervised learning approaches and need tons of labeled data to excel.

Label acquisition for 3D Computer Vision tasks and specifically for the human mesh estimation task became even more difficult and error-prone compared to 2D tasks due to the inherent complexity of working with an additional dimension. That is where the synthetic data steps in, allowing researchers to obtain much more cost-efficient and pixel-perfect annotations.

In this work, we utilize the existing Latent Diffusion Model for conditional image generation and create a method for synthesizing a large dataset of humans with 3D mesh labels obtained without the involvement of a human annotator. Further, we show the effectiveness of using such a synthetic dataset and its superiority compared to other synthetic data obtained from the game engines. The implementation of the proposed approach can be accessed on the GitHub[1].

[1] https://github.com/viniavskyi-ostap/synth-smplerx

Acknowledgements

First of all, I want to thank my supervisor, Orest Kupyn, for valuable advice and guidance throughout the work on this thesis. Also, I am grateful to Bohdan Hlovatskyi for the many discussions we had on the topic and the insights it brought. Many thanks to the Ampersand Foundation for believing in me by providing the scholarship for my Master's studies and making this work possible.

Contents

List of Figures

List of Tables

List of Abbreviations

SMPL	**S**kinned **M**ulti-**P**erson **L**inear (Model)
SMPL-X	**S**kinned **M**ulti-**P**erson **L**inear e**X**pressive (Model)
ViT	**V**ision **T**ransformer
CLIP	**C**ontrastive **L**anguage-**I**mage **P**re-Training
GAN	**G**enerative **A**dversatial **N**etwork
VAE	**V**ariational **A**uto-**E**ncoder
LDM	**L**atent **D**iffusion **M**odel
DDPM	**D**enoising **D**iffusion **P**robabilistic **M**odel
DDIM	**D**enoising **D**iffusion **I**mplicit **M**odel
IP-Adapter	**I**mage **P**rompt Adapter
T2I-Adapter	**T**ext-**to**-**I**mage Adapter
LoRa	**Lo**w-**Ra**nk (Approximation)
IMU	**I**neartial **M**easurment **U**nit

Dedicated to my family

Chapter 1

Introduction

1.1 Motivation

Methods that try to detect and describe humans in the images are central to the Computer Vision field. Human recognition tasks in Computer Vision range from bounding box detection, 2D pose estimation, and semantic segmentation that try to understand the position and semantics of humans in the image to more complicated tasks like 3D detection, 3D pose estimation, and recovery of human avatars that try to uncover 3D geometry of human body in the image.

A significant part of the success of the methods that solve these tasks is attributed to large labeled datasets used to train supervised models. One such example can be the Microsoft COCO dataset [Lin et al., 2015] introduction, which led to considerable success in the 2D human pose estimation task. Yet, obtaining the large labeled datasets is usually quite expensive and time-consuming. It can only be affordable for large companies or labs, limiting their abundance in more niche tasks.

An even bigger problem persists for 3D understanding tasks since labeling of 3D data is even more time-consuming and error-prone. For instance, the task of 3D pose estimation leaves a lot of ambiguity for human annotators, especially if only a single view of the labeling target is available. Another example can be the labeling of the DensePose dataset [Ho, Jain, and Abbeel, 2020], which established mapping from the human body's pixels to the prepared in advance avatar. The labeling process was highly sophisticated and, in the end, produced only very sparse annotations. This shows the need for another alternative solution for labeling the images with 3D annotations, particularly images of humans.

Significant progress in the field of 3D vision was achieved thanks to synthetic data. The promising approach for collecting large 3D annotated datasets is utilizing advances in the graphics pipeline and creating synthetic humans from the renders of human 3D models. These approaches [Patel et al., 2021a, Black et al., 2023a Erfanian Ebadi et al., 2022, Yan et al., 2021] allow the collection of complete information about the 3D data present in the image from the human pose and shape to pixel-wise mapping from the image to the surface of the human mesh. The major downside is the quality of the renders, the complexity of creating human models in diverse scenarios, and the complexity of clothing human avatars.

Deep Generative models have quite a long history of usage as a tool for synthesizing additional training data since the success of Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] or generating realistic augmentations for existing data. Nevertheless, the human generation task was not feasible until recently. The Denoising Diffusion Probabilistic (DDPM) [Ho, Jain, and Abbeel, 2020] models pushed the boundary of what is possible for human generation, with large-scale models like Stable Diffusion [Rombach et al., 2022] capable of generating realistic humans in diverse poses, shapes, clothing, lighting, etc.

In this work, we describe the possible applications of Deep Generative models for generating diverse datasets of humans with annotations provided by the conditional generation process. We propose the approach for generating humans with various poses, shapes, and clothing and use the generated datasets to train human pose and shape recognition models.

1.2 Contributions

The contributions of this work are the following:

- We explore the limitations of current datasets, both real-world and synthetic, that are used for training human pose and shape estimation models.
- We propose a controllable human generation approach that preserves the distribution of real-world images and, at the same time, allows the collection of 3D labels. The method is fully automatic and does not require any involvement on the annotator side.
- We solve the downstream task of human mesh estimation using only our synthetic data and evaluate results compared to methods that use other real-world and synthetic data.
- We show that our synthetic dataset achieves comparable or better results than the methods trained on other synthetic datasets obtained from the 3D game engines. This showcases that expensive synthetic human data acquisition through modeling humans in the game engines can be replaced by the cheap generation using Deep Generative models.

1.3 Structure Of The Thesis

The thesis is structured as follows. In Chapter 2, we introduce the task of human pose and shape estimation and describe the SMPL-family [Loper et al., 2015] of parametric 3D human models. We also describe the theoretical basis of diffusion models and the possibilities for introducing control in the generation process. In Chapter 3, we review the related works, discussing the existing datasets for human recognition tasks and how they can be utilized for the human mesh estimation task. We also review existing approaches that generate humans and utilize synthetic data for specific tasks. We describe the downstream task of particular interest to us - human pose and shape estimation. In Chapter 4, we describe our methodology for controllable human generation in detail, while in Chapter 5, we provide the details of experiments and quantitative and qualitative results. Finally, we conclude our work and propose directions for future research in Chapter 6.

Chapter 2

Theoretical background

The recognition of people in images is a long-standing problem in the area of computer vision. Simpler tasks aim at estimating the position of the human in the image or the skeletal structure of the human body projected onto the image. These are tasks of human detection and 2D human joint estimation, respectively.

More sophisticated approaches try to estimate the 2D projections of human joints and the 3D position of a human skeleton or even recover a full 3D mesh of the human body surface in the image, in other words, a human avatar. At first sight, estimating a 3D structure from a single image might seem like an ill-posed problem. However, due to the constrained nature of the human body, this can be done quite successfully [Baradel et al., 2024].

2.1 Parametric human body models

Representing the human body surface as a free-form 3D mesh offers benefits such as almost unlimited expressivity, allowing us to capture the smallest details of the human body. Yet, such representation is hard to estimate and maintain if we, for example, want to track the movement of the human body in time.

The 3D mesh with 10,000 3D vertices and predefined topology has 30,000 degrees of freedom, which must be calculated during the estimation stage or stored independently for each time step. Also, the slight movement of the human body will require recalculating a large portion of vertices locations, which further complicates working with such a representation.

Since the locations of many vertices are correlated, we can derive an underlying representation of a much smaller size that will effectively capture the distribution of possible human bodies. Also, we need to define a transformation from such representation to the 3D mesh to be able to recover the full mesh effectively.

In this work, we choose to work with the SMPL [Loper et al., 2015] parametric human body model and its successor representation SMPL-X [Pavlakos et al., 2019], which are de facto standards for the parametric human body modeling. SMPL decomposes the parameters into two sets: pose and shape, which control the pose of the human skeleton and the shape of the human body independent of the current pose. SMPL-X adds to the previous model by making expressive hands and faces. This is achieved by including fingers in the underlying skeleton and adding additional parameters to control facial expression.

2.1.1 SMPL and SMPL-X models

The Skinned Multi-Person Linear model (SMPL) is created to represent the human bodies of various shapes and poses using a small set of parameters. The main advantage of SMPL is that it decomposes the shape and pose, allowing one to control

both independently. Also, it accounts for the shape deformations resulting from the changes to the pose.

SMPL defines a single mesh topology consisting of $N = 6890$ vertices and $F = 10475$ faces connecting the vertices. It also defines an underlying skeleton consisting of $K = 24$ joints. Pose parameters $\vec{\theta} \in \mathbb{R}^{3K}$ stand for three rotation parameters in each skeleton joint, where rotation of the root joint (pelvis) defines the global rotation of mesh with respect to the camera, and other joints store rotations relative to the parent joint in the kinematic tree. Positions of the mesh vertices are derived via the Linear Blend Skinning (LBS) process, which calculates the rotation of each vertex as a linear combination of rotations of predefined joints for this particular vertex.

The shape is controlled by the parameters $\vec{\beta} \in \mathbb{R}^S$, which add shape-dependent correctives to the template mesh. The shape parametrization was derived through the PCA [Shlens, 2014] on the large collection of human meshes normalized to the neutral pose. The principal components define the so-called shape blendshapes, while coefficients near those principal components are our parameters $\vec{\beta}$. S is chosen to be up to 300 since such a number of principal components is enough to capture most of the variance in human shapes. In practice, we use the reduced representation, where $S = 10$. Also, SMPL defines pose blendshapes, which add shape correctives dependent on the human pose. This is needed to alleviate the negative effects of the LBS process, such as the candy wrapper problem.

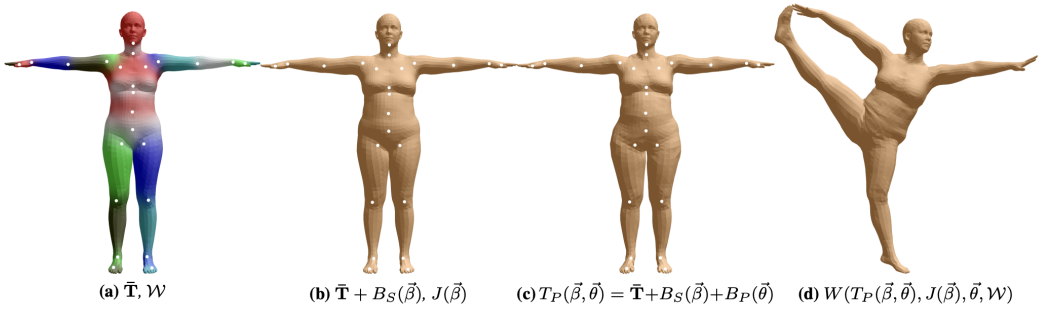

FIGURE 2.1: SMPL mapping stages. Taken from [Loper et al., 2015]

More formally, SMPL defines a mapping from pose and shape parameters to the final posed and shaped mesh $M(\vec{\beta}, \vec{\theta}; \Phi) : \mathbb{R}^{|\vec{\theta}| \times |\vec{\beta}|} \mapsto \mathbb{R}^{3N}$, where Φ are constants learned from the data. Φ includes several distinct entities: $\bar{\mathbf{T}}$ - template mesh in normalized, so-called T-pose, and mean shape, $\mathcal{W} \in \mathbb{R}^{N \times K}$ - sparse LBS weights matrix that defines the influence of joints rotations on the vertex rotations, $J(\vec{\beta}) : \mathbb{R}^{|\vec{\beta}|} \mapsto \mathbb{R}^{3K}$ - regression parameters for inferring locations of joints from shape parameters, $B_S(\vec{\beta}) : \mathbb{R}^{|\vec{\beta}|} \mapsto \mathbb{R}^{3N}$ - shape blend shape parameters, $B_P(\vec{\theta}) : \mathbb{R}^{|\vec{\theta}|} \mapsto \mathbb{R}^{3N}$ - pose blendshapes.

The SMPL mapping can be broken down into several distinct stages as depicted in Figure 2.1. We start with the template mesh $\bar{\mathbf{T}}$ and add pose blendshapes to it to obtain shaped mesh $T_S(\vec{\beta})$

$$T_S(\vec{\beta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) \tag{2.1}$$

From this shaped mesh, we regress joint locations $\mathbf{J}$, effectively making the joints dependent only on the shape parameters.

Next, we add pose-dependent correctives to the shaped mesh

$$T_P(\vec{\beta}, \vec{\theta}) = T_S(\vec{\beta}) + B_P(\vec{\theta}) = \bar{\mathbf{T}} + B_S(\vec{\beta}) + B_P(\vec{\theta}) \tag{2.2}$$

Using rotation parameters, we first deform skeleton joint locations to the target pose, and then through the process of LBS, we obtain the final locations of the posed and shape mesh:

$$M(\vec{\beta},\vec{\theta}) = W(T_P(\vec{\beta},\vec{\theta}),\mathbf{J}(\vec{\beta}),\vec{\theta},\mathcal{W}) \tag{2.3}$$

Individual vertex location is computed as follows:

$$\mathbf{t}'_i = \sum_{k=1}^{K} w_{k,i} G'_k(\vec{\theta},\mathbf{J})(\bar{\mathbf{t}}_i + \mathbf{b}_{S,i}(\vec{\beta}) + \mathbf{b}_{P,i}(\vec{\theta})) \tag{2.4}$$

$$G'_k(\vec{\theta},\mathbf{J}) = G_k(\vec{\theta},\mathbf{J})G_k(\vec{\theta}^*,\mathbf{J})^{-1} \tag{2.5}$$

$$G_k(\vec{\theta},\mathbf{J}) = \prod_{j\in A(k)} \left[\begin{array}{c|c} \exp(\vec{\omega}_j) & \mathbf{j}_j \\ \hline \vec{0} & 1 \end{array}\right], \tag{2.6}$$

where $G_k(\vec{\theta},\mathbf{J})$ and $G'_k(\vec{\theta},\mathbf{J})$ are transformations of individual joints with respect to the world origin and parent joint location respectively, $j \in A(k)$ - child joints for parent joint k, $\exp(\vec{\omega}_i)$ - mapping from the axis-angle rotation representation to the rotation matrix.

SMPL-X model adds to the SMPL model by introducing facial expression parameters $\vec{\psi} \in \mathbb{R}^{10}$, which are derived analogously to the shape parameters for the human body and are accompanied by the face expression blendshapes. Also, SMPL-X extends the skeleton with 30 new joints: 15 for each hand, allowing control of the pose of individual fingers. 90 new rotation parameters are often redundant to control the hand pose. Thus, the hand pose is further coded via PCA into lower dimensional space. In all other ways, SMPL-X is analogous to the original SMPL model.

2.1.2 Limitations of SMPL

The SMPL model is derived to represent various human poses and shapes by utilizing only a relatively small set of disentangled pose and shape parameters. Yet, because of its constrained nature, the SMPL is limited in representing all possible poses and shapes.

Moreover, regressing SMPL parameters from the image is a complex task. SMPL representation is inherently a 3D representation; thus, to estimate the correct mesh parameters for the person on the image, we also need to possess the correct camera information to be able to project the derived mesh onto the image accurately. Most of the existing state-of-the-art approaches [Zhang et al., 2021, Cai et al., 2023] assume a simplified camera model by using the weak-perspective camera or setting the focal length of the perspective camera to a large constant value. This results in predictions that are quite accurate in the 3D pose while having poor alignment of the projected avatar with the true human silhouette. This limitation will become important when we try to learn a controllable human generator from predefined SMPL avatars. The problem can be seen in the Figure 2.2

SMPL limitations also include entangled shape parameters and the possibility of generating invalid avatars from the valid input parameters. The former means that individual shape parameters do not influence a particular part of the body but can change the whole avatar simultaneously. For example, if we want to derive from the current avatar mesh with the same shape except for waist width, we will likely need to change all the shape parameters at once. The latter means that not all SMPL parameters lead to the valid meshed of humans. For instance, some rotation of joints

FIGURE 2.2: Silhouette alignment problem of mesh predicted by SMPLer-X [Cai et al., 2023]. Input image (right), color-coded render of mesh (middle), overlay of image and mesh (right)

can lead to unrealistic bending of human parts, or the overall parameter setting can lead to human meshing with self-collisions. Thus, sampling of SMPL parameters cannot be done uniformly over the parameter space; rather, it needs more elegant approaches.

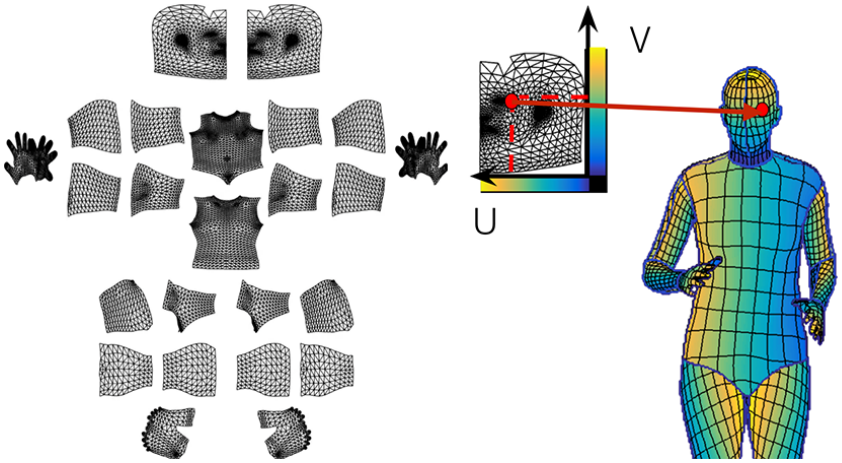

FIGURE 2.3: DensePose Representation. The template mesh is split into multiple parts, and unwrap UV coordinates are defined for each body part. Taken from [Güler, Neverova, and Kokkinos, 2018]

2.1.3 DensePose

DensePose [Güler, Neverova, and Kokkinos, 2018] is a 2.5D representation connecting 2D images and 3D human meshed, particularly the SMPL model. DensePose does not directly represent the human in the image with the SMPL parameters and corresponding camera to project the constructed avatar. In contrast, it provides a binary silhouette mask of the human and defines a mapping from each human pixel to the template mesh, which is the SMPL template mesh. Moreover, DensePose splits

the avatar into several parts and defines for each image pixel to which body part it belongs and where it is located specifically on that body part. The latter uses UV coordinates of the UV unwrap of the particular body part from the template mesh. The process is visualized in the Figure 2.3.

Also, the authors train the model to predict the body-part index (I) and corresponding coordinates in the unwrapped space (UV), creating a DensePose-IUV representation of the human in the image. The model was trained on manually collected correspondences between images and template mesh through a tedious labeling process. The released model has quite accurate predictions of body part index but lacks accuracy in predicting UV maps. Thus, the DensePose index is usually the only thing from two used for human representations [Chang et al., 2023].

The main advantage of DensePose is that we can have pixel-accurate human representation predicted on the in-the-wild images of humans. At the same time, we can easily convert any SMPL human mesh to the DensePose render, which will be useful later when we want to generate humans in a controllable way. The difference between the DensePose part index and the DensePose rendered from the estimated avatar is shown in Figure 2.4.

FIGURE 2.4: DensePose representation. From left to right: input image, predicted DensePose part index, overlay of DensePose with image, overlay of estimated avatar and image, estimated avatar rendered to the DensePose part index

2.2 Conditional Image Generation

This section explains the basics of controllable or conditional image generation. Most of the approaches utilize recent progress made around Denoising Diffusion Probabilistic Models [Ho, Jain, and Abbeel, 2020] and Latent Diffusion Models [Rombach et al., 2022].

2.2.1 Diffusion Models

Denoising Diffusion Probabilistic Models [Ho, Jain, and Abbeel, 2020] revolutionized the field of generative modeling of images. They allow for generating images of higher fidelity compared to Generative Adversarial Networks [Goodfellow et al., 2014] and don't suffer from the mode collapse problem and unstable training common for GANs.

Diffusion Models define two processes: forward and backward. In the forward process, the input data, image in our case, is noised by adding random Gaussian noise according to the predefined schedule $\beta_1, \ldots, \beta_T$. More formally, the forward

noising process for the initial data point $\mathbf{x}_0$ is defined as follows:

$$q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right), \quad q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \tag{2.7}$$

The reverse process starts with a sample from standard Normal distribution $p\left(\mathbf{x}_T\right) = \mathcal{N}\left(\mathbf{x}_T; \mathbf{0}, \mathbf{I}\right)$ and transforms noise into data point by the learned Gaussian transition. Again, the reverse process is defined as:

$$p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right), \quad p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right), \tag{2.8}$$

where θ are the learnable parameters of the network.

The model parameters are optimized by minimizing the VLB on the negative log-likelihood:

$$\begin{aligned} \mathbb{E}\left[-\log p_\theta\left(\mathbf{x}_0\right)\right] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta\left(\mathbf{x}_{0:T}\right)}{q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right)}\right] \\ &= \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right) - \sum_{t \geq 1} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right] =: L \end{aligned} \tag{2.9}$$

2.3 Latent diffusion models

Learning the diffusion model in the image space is prohibitively slow for high-resolution images. Authors of the Latent Diffusion Model [Rombach et al., 2022] proposed mapping the input images to the compressed latent space of decreased resolution and then learning the diffusion model over the distribution of images in the latent space. The mapping to the latent space is performed by the Variational Autoencoder (VAE) [Kingma and Welling, 2019], which was trained to preserve local detail in the image. In this way, VAE is responsible for compressing low-level details in the image. At the same time, the Diffusion model learns semantics and object composition in the image, which it is best suited for. The prominent architecture from the family of Latent Diffusion Models is Stable Diffusion, which we utilize in our work.

2.4 Conditioning Diffusion Models

With the introduction of diffusion models, multiple methods for their conditioning were introduced. Classifier-guided diffusion models [Dhariwal and Nichol, 2021] allowed for higher fidelity generation by utilizing a separate classifier model, which was used during the inference process to steer the denoising path into the direction that maximized the likelihood of one of the classes. This allowed to have two-fold advantages: class label conditioning and improved generation quality with the tradeoff of decreased variability. On the contrary, classifier-free guidance [Ho and Salimans, 2022] removed the necessity of training a separate classifier by introducing free-form text conditioning into the diffusion model.

Text conditioning was introduced using a pretrained CLIP model [Radford et al., 2021], which greatly aligns text features with visual ones. Stable Diffusion model

utilizes cross-attention [Vaswani et al., 2023] mechanism to introduce CLIP text features into the denoising model.

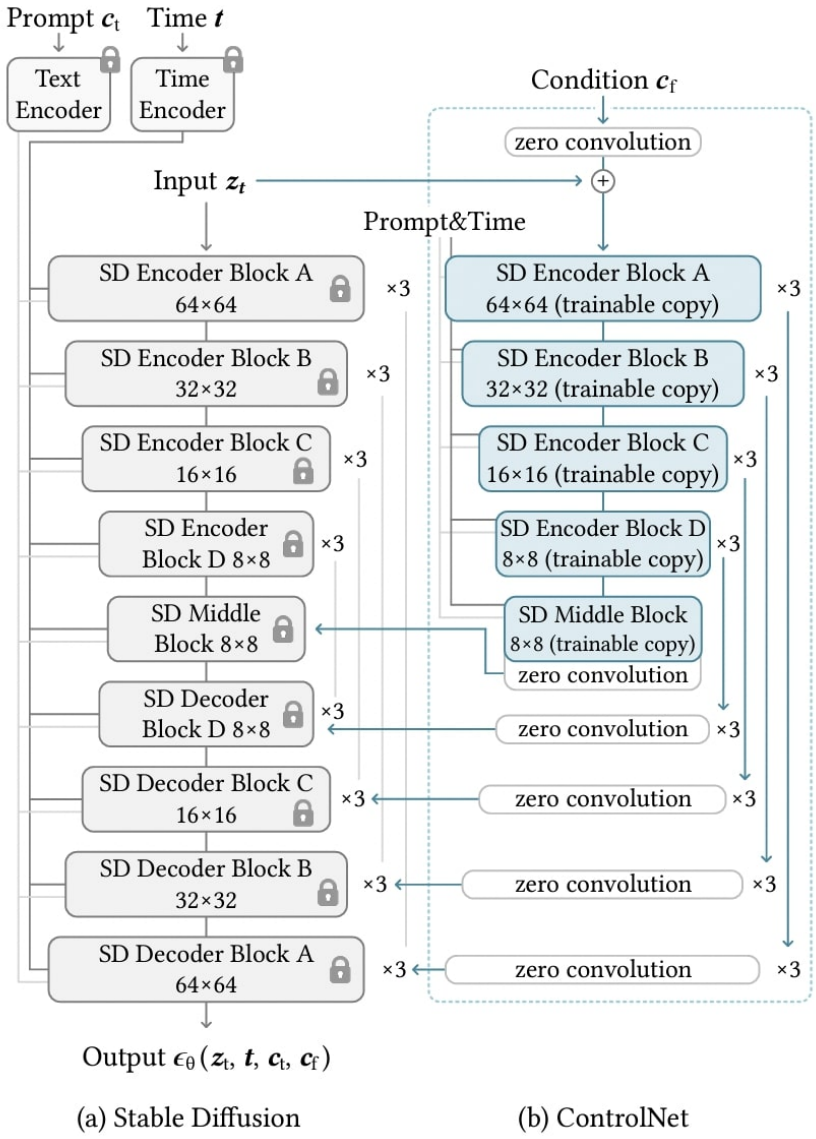

FIGURE 2.5: ControlNet architecture. Spatial features from the control modality are added pointwise with the spatial features of denoising UNet in the decoder. Taken from [Zhang, Rao, and Agrawala, 2023]

2.4.1 Image Conditioning

Introducing the image as a condition to the generation process can be done similarly to the text condition. IP-Adapter [Ye et al., 2023] was the first one to achieve great results with such an approach without retraining the foundational text-to-image Stable Diffusion model. They encoded conditional images with the CLIP vision model and then transformed the features into tokens of the same dimensionality as text tokens in the original text conditioning. Then, they trained only the cross-attention layer that infused conditional image features into the denoising model.

IP-Adapter is useful for capturing the style and rough context of the input conditional image but cannot control the precise location and appearance of the generated scene due to the highly compressed nature of the CLIP image features.

2.4.2 Adding spatial control to the denoising process

In contrast to the IP-Adapter, such methods as ControlNet [Zhang, Rao, and Agrawala, 2023] and T2I-adapter [Mou et al., 2023] took different mechanisms to introduce conditions into the denoising UNet model [Ronneberger, Fischer, and Brox, 2015]. They propagate spatial features into the UNet skip connections, making them effective at aligning generated content with a given control. The mechanism of ControlNet introducing features into the Stable Diffusion is visualized in Figure 2.5. They also are adapter methods, which means they don't require retraining the full diffusion model from scratch and usually can be combined with other adapters.

2.4.3 Other conditioning possibilities

There exist multiple other possibilities for introducing conditions into the Diffusion Model that we introduce here. The most naive one is concatenating the conditioning modality to the noise image that is fed to the denoising model. In this way, StableDiffusion [Rombach et al., 2022] achieves depth-conditional generation and inpainting capabilities for the model. The inpainting in this way can be seen as a conditional generation, where only part of the image is visible and is provided as a condition. In such a setup, the model requires full fine-tuning, requiring enormous resources and a time budget for each new conditioning modality.

One of the solutions to alleviate this problem is using LoRAs [Hu et al., 2021] to fine-tune only low-rank approximation matrices, which are added to the weights. Such an approach allows for decreasing computation and memory requirements for the model training at the cost of decreased controllability and model expressivity. Also, mixing multiple conditions with different LoRAs will not be as trivial as it is with adapter-based methods.

Considering the above, we do not further examine fine-tuning-based methods for conditional generation with DDPM; instead, we rely solely on adapter-based approaches. On the other hand, in this work, we explore different modalities that might be useful for generating images of humans in a controllable manner.

Chapter 3

Related works

In this chapter, we explore the works related to this study. The chapter will be divided into three parts. Firstly, we start with an overview of the existing real-world and synthetic datasets for human recognition tasks, specifically for human pose and shape estimation. We identify their strong and weak points, which can be later alleviated in our process of creating a synthetic dataset. Next, we proceed with the overview of the methods that generate synthetic humans similar to ours. We finish with the exploration of the human avatar estimation method that we will use as a basis for our experiments.

3.1 Existing datasets overview

Learning human mesh estimation models is a complex task that requires lots of training data. In the most naive scenario, one will directly predict SMPL [Loper et al., 2015] parameters of the human avatar from the image and optimize the model toward predicting ground truth parameters. This requires the knowledge of the true parameters, which are hard to obtain for real-world data. The other alternatives are to utilize existing datasets with ground truth 2D or 3D joints labeled or other visual clues like virtual markers [Ma et al., 2023]. In such a process, the loss will be calculated between visual clues of predicted and projected, if needed, avatar and the true visual clues. The gradients will be propagated to the SMPL parameters of the predicted avatar through the processes of avatar creation and camera projection, which are fully differentiable.

3.1.1 Real-world datasets

Multiple datasets exist that provide 3D annotations for images of humans, but they come with several limitations. Many methods use 3D information acquisition devices like RDBG cameras, LIDAR scanners, or inertial measurement units (IMU) to obtain 3D geometry of the scene and, specifically, the poses and shapes of humans. For instance, the Human3.6m [Ionescu et al., 2014] dataset consists of 3.6 million human poses captured by the motion capture system. The dataset is captured in a constrained environment with 11 actors acting in one of seventeen scenarios. The 3DPW [Marcard et al., 2018] provides in-the-wild videos with 3D human skeletons annotated from both video and IMU attached to the human. With 60 video sequences, the diversity of the data is still minimal. The MPI-INF-3DHP [Mehta et al., 2017] provides another 1.3 million frames from 14 cameras simultaneously capturing actors in the lab environments. The above datasets provide only 3D human skeletons, thus containing minimal information about the body shape. Also, their capturing methodology restricts them to very limited scenarios, clothing types, and lighting conditions.

FIGURE 3.1: Rendered images (top row) from AGORA dataset [Patel et al., 2021a] and corresponding overlay of avatars (bottom row)

Several datasets provide full avatar reconstruction of the human body obtained either from complex and expensive capturing systems - CAPE [Ma et al., 2020] or EHF [Pavlakos et al., 2019] and have the same pitfalls as the datasets with 3D pose, or use pseudo ground truth data optimized from visual cues and often suffer from the severe inaccuracies. An example of the latter is the NeuralAnnot [Moon, Choi, and Lee, 2022, Moon et al., 2023], which labels avatars from the ground truth annotations like 2D and 3D joints in the other datasets.

3.1.2 Synthetic data from 3D engines

The promising approach for collecting large 3D annotated datasets is utilizing advances in the graphics pipeline and creating synthetic humans from the renders of human 3D models. These approaches allow the collection of complete information about the 3D data present in the image, from the human pose and shape to pixel-wise mapping from the image to the surface of the human mesh. The significant downsides are the quality of the renders, the complexity of creating human models in diverse scenarios, and the complexity of clothing human avatars.

Some available datasets, which we collect from commonly used academic benchmarks, like AGORA [Patel et al., 2021a], Surreal [Patel et al., 2021a], PeopleSansPeople [Erfanian Ebadi et al., 2022] provide hundreds of thousands of fully annotated images, but suffer from the problems mentioned above with quality and diversity. Ultrapose [Yan et al., 2021] utilizes commercial software for generating higher quality images, capturing 1 Billion points mapping from the images to 3D avatars for the task of DensePose estimation, but does not release their dataset. Examples of scenes with humans and corresponding true avatars from the AGORA dataset can be seen in Figure 3.1.

3.2 Synthetic human generation

High-fidelity human generation through generative modeling was not possible until the introduction of StyleGAN Human [Fu et al., 2022] based on StyleGAN-V2 [Karras et al., 2020] architecture for unconditional human generation. The method can generate clothed humans, but the versatility of poses is quite limited, and the generation quality is relatively low, often introducing strong visible artifacts such as deformed faces and hands.

Introduction of text-to-image Latent Diffusion Models [Rombach et al., 2022, Podell et al., 2023] trained on a large dataset of internet images LAION [Black et

FIGURE 3.2: Text prompt from the FIRST approach [Huang et al., 2023] (top row) and corresponding generated image (bottom row)

al., 2023a] paired with the text description enabled high-quality generation of versatile objects and scenes including humans. Synthetic data generated through conditional generation utilizing diffusion models was shown to boost the performance of the models in several tasks. For example, [Azizi et al., 2023, Sariyildiz et al., 2023, Shipard et al., 2023, Tian et al., 2023] show how synthetic data can improve the performance of the classification models, but are limited to only this problem.

The conditional generation of humans is also utilized to advance other tasks through the generation of synthetic data. Recent method SewFormer [Liu et al., 2023] used text-to-image translation to improve the quality of their rendered dataset for sewing pattern prediction. In this way, they are able to obtain a dataset with several millions of entries and train a model that achieves state-of-the-art in the pursued task. On the other hand, FIRST [Huang et al., 2023] introduced a million-entry synthetic dataset for text-driven fashion synthesis, allowing for a highly controlled human generation process through detailed text descriptions. Yet, many essential tasks like DensePose prediction or human 3D pose and shape estimation still lack the attention of researchers and are bottlenecked by relatively small and not diverse datasets.

3.3 SMPLer-X

Lastly, we review the state-of-the-art method for human SMPL-X mesh estimation called SMPLer-X [Cai et al., 2023]. We chose this method for the experiments part since it establishes a foundational model for the single human avatar estimation task. The method unifies training on multiple datasets, both real-world and synthetic, with different ground truth labels that include 2D and 3D joints and true SMPL-X avatar parameters. This establishes a convenient framework for testing multiple backbone models, providing different visual clues as ground truth data, as well as comparing the performance of our synthetic dataset against other datasets often used in the community.

SMPLer-X introduces a simple architecture based on the ViT backbone [Dosovitskiy et al., 2021]. The image is split into multiple patches that are flattened into the tokens. Additionally, learnable task tokens are concatenated to the input, representing pose, shape, and camera prediction tokens. Significantly, SMPLer-X does not directly predict SMPL-X parameters but first predicts the person's root-relative 3D joints, which are then sent to the avatar parameters prediction head. Furthermore,

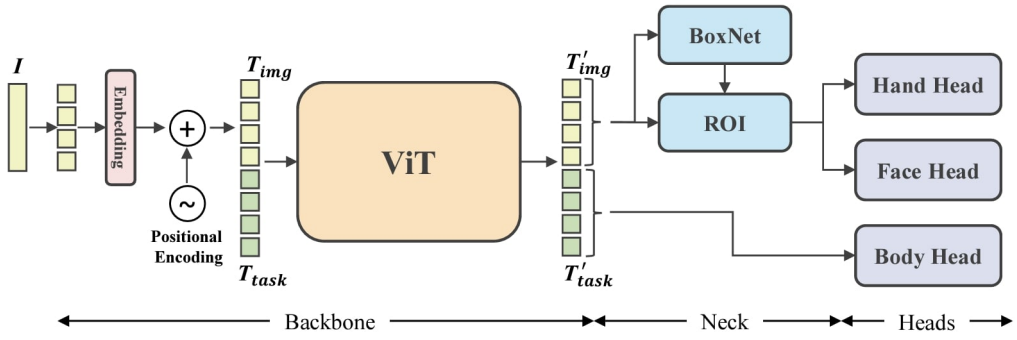

FIGURE 3.3: SMPLer-X architecture. Taken from [Cai et al., 2023]

separate branches are utilized to predict the hands and face joints and landmarks. The high-level architecture of the SMPLer-X model can be seen in Figure 3.3.

SMPLer-X uses several loss functions to learn the model parameters. Firstly, it introduces loss in the SMPL-X parameters space for pose, shape, and expression parameters separately:

$$\begin{aligned} \mathcal{L}_{pose} &= \sum_i |\theta_i - \hat{\theta}_i| \\ \mathcal{L}_{shape} &= \sum_i |\beta - \hat{\beta}_i| \\ \mathcal{L}_{expr} &= \sum_i |\psi - \hat{\psi}_i|, \end{aligned} \tag{3.1}$$

where θ, β, ψ are ground truth and $\hat{\theta}$, $\hat{\beta}$, $\hat{\psi}$ are predicted pose, shape and expression parameters respectively. SMPLer-X also introduced loss function over the predicted joints both in 3D and 2D.

$$\begin{aligned} \mathcal{L}_{3D} &= \sum_i ||j_i^{3D} - \hat{j}_i^{3D}||_1 \\ \mathcal{L}_{2D} &= \sum_i ||j_i^{2D} - \hat{j}_i^{2D}||_1, \end{aligned} \tag{3.2}$$

where j^{3D}, j^{2D} are ground truth and $\hat{j}^{3D}$, $\hat{j}^{2D}$ are predicted 3D and 2D joints respectively.

The joint loss function is calculated twice over the initial joint predictions and the joints extracted from the built SMPL-X avatar.

The main contribution of SMPLer-X is unifying training on multiple datasets, some of which have only 2D or 3D ground truth joints available. In such cases, parameter loss functions are deactivated, and only joint loss functions are used for training. SMPLer-X provides a comprehensive study of the quality of the datasets used for the avatar estimation problem, as well as the scaling laws for the data points used for training.

Unifying tens of datasets into one training is a tedious process, which requires lots of data preparations and loss function balancing in the training phase. We hypothesize that we can achieve results comparable to the SMPLer-X model with much

simpler training on purely synthetic data, where all ground truth data is always available.

3.4 Motivation for synthetic dataset creation

To sum up this chapter, we state the need for the creation of another synthetic dataset for the human mesh recovery task. Current real-world datasets provide either incomplete annotations for human avatar estimation or obtain their labels from the sub-optimal optimization process. Existing synthetic datasets provide full annotations and images perfectly aligned with them. Yet, they are expensive to obtain and severely lack diversity.

At the same time, in-the-wild human pose and shape estimation from a single image is a highly ambiguous task. Different 3D poses project to the similar 2D poses in the image. Moreover, human bodies are often occluded by the wide clothing, other people, or simply other objects. Thus, we need a dataset that simultaneously provides complete labels, meaning that ground truth 3D human mesh is available and is diverse and realistic enough to capture the real-world distribution of human images. In this work, we show that such a dataset can be created with the means of conditional generation utilizing Latent Diffusion Models.

Chapter 4

Method

In this chapter, we describe the proposed approach for a controllable human generation with the intent of creating a synthetic labeled dataset for the task of human mesh estimation. We start by explaining the chosen conditional image generation process and move to the description of the label sampling process. Finally, we conclude the chapter with the method for assessing the quality of the synthetically created dataset.

More formally, our goal is to generate a dataset D:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\},$$

consisting of m image-label pairs where labels are ground truth SMPL-X avatar parameters. Most human mesh estimation methods [Cai et al., 2023] follow a top-down approach, where first, humans are detected in the image with the off-the-shelf detector. Then, avatar estimation is done on the tight crop around the detected person. In this work, we follow this paradigm and focus only on a single human generation, as such images will be enough to train a mesh estimation model that works in a top-down manner.

Given a conditional generative model $G(y)$ capable of generating images x_i from a condition y_i, we start with sampling the labels $y_i \sim Y$ from some label distribution Y. Assuming that the produced image x_i follows the y_i, we can generate a dataset D suitable for supervised learning on the downstream task. As our base generative model, we choose the Stable Diffusion 1.5 [Rombach et al., 2022] model, which allows for high-quality controllability. Specifically, we use the Realistic Vision v5.1 [SG_161222, 2023] checkpoint of Stable Diffusion, which was fine-tuned for the realistic human generation.

4.1 Conditional image generation

First, we develop a methodology for controllable human generation from the SMPL-X avatar using the adapter-based method of ControlNet [Zhang, Rao, and Agrawala, 2023]. We chose ControlNet for its ability to add spatial control to the generation process in such a way that full retraining of the model is not required. Also, ControlNet does not require large quantities of training data as is used in the training of the base model. Usually, a few hundred thousand image-control pairs are enough to achieve great results.

ControlNet is a method for spatial conditioning, meaning that it is excellent at controlling the location and shape of the desired objects in the generated image. This results from its conditioning mechanism, where spatial feature maps from ControlNet are added point-wise to the spatial feature maps in the denoising UNet decoder. Thus, to achieve conditioning from the SMPL-X human avatar, we need to render

it to a 2D image. The rendering should also preserve some information about the human's 3D pose in the image.

The typical ControlNet training pipeline is straightforward in its nature. Given a dataset of unlabeled images, we predict the conditional modality using an off-the-shelf predictor and use it as a control input to the ControlNet. This approach works well for such modalities as 2D human pose, semantic segmentation, or edges in the image because state-of-the-art predictors provide estimates that are well-aligned with the visual clues in the input image.

We follow such a training paradigm for SMPL-X avatars, where we utilize the state-of-the-art avatar predictor for control estimation and show that it is sub-optimal to do so. Next, we rely on the intermediate representation of the DensePose for ControlNet training and show its superiority compared to the training on the direct avatar estimates.

4.1.1 Color-coded rendering

We start with the naive ControlNet training approach. Given an unlabeled human images dataset, we label it with a state-of-the-art avatar predictor SMPLer-X to form image-control pairs. To preserve the information about human pose after the rendering, we color-code the avatar vertices with unique RGB colors that change smoothly over the avatar surface. For this purpose, we directly map the X, Y, and Z coordinates of the avatar vertices in the mean pose to the red, green, and blue channels of RGB color. We call this representation a color-coded avatar render or CC-avatar for short.

Such an approach does not provide a satisfactory result. As mentioned in Chapter 2, human mesh estimation models suffer from a poor alignment problem with the true human silhouette in the image. This results in the effect that our dataset for ControlNet training has avatar renders that are quite severely misaligned with the humans in the image. Consequently, ControlNet training becomes unstable, and the resulting adapter model produces poor-quality final generations.

An example of training pairs for ControlNet with CC-avatars as an input condition can be seen in the first two columns of Figure 4.1. The image generated with such a ControlNet model can be seen in the third column. It has severe misalignments, especially in the shoulders, head, and arms of the human body.

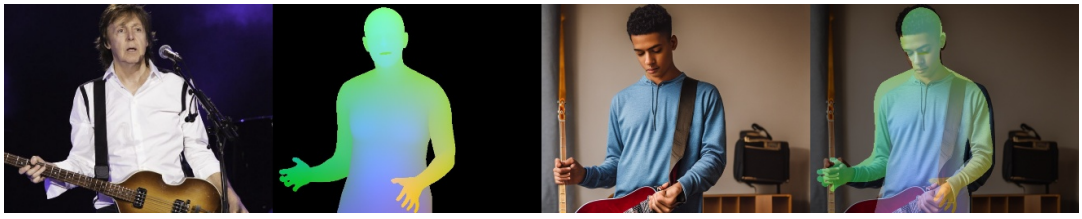

FIGURE 4.1: Conditional image generation from CC-avatars. Left to right: original image, estimated CC-avatar, generation from estimated CC-avatar, overlay of the generated image with its control

4.1.2 DensePose intermediate representation

DensePose can be used as an intermediate representation for building a conditional human generation model. Training ControlNet from the DensePose modality is

FIGURE 4.2: Conditional image generation from DensePose. Left to right: original image, estimated DensePose using [Güler, Neverova, and Kokkinos, 2018], generation from estimated DensePose, overlay of the generated image with its control

FIGURE 4.3: Conditional image generation from avatar rendered to DensePose. Left to right: original image, estimated SMPL-X avatar rendered to DensePose, generation from avatar DensePose, overlay of the generated image with its control

much more appealing than from color-coded avatar render since there exists a predictor [Güler, Neverova, and Kokkinos, 2018] that can accurately estimate the DensePose part index. This predictor was trained on manually labeled correspondences between the human body in the image and template mesh. Since no avatar estimation is involved in predicting DensePose, such a part index map is well aligned with the true silhouette of the human in the image.

The power of the DensePose-based ControlNet is that it can generate images both from DensePose predicted directly by the neural network [Güler, Neverova, and Kokkinos, 2018], as well as from 3D human mesh. This is because any SMPL or SMPL-X avatar can be rendered into the DensePose representation, as the mapping of mesh vertices to different body parts is completely pre-defined. The reverse statement is not true since the mapping from the predicted DensePose to the SMPL-X avatar parameters is not trivial. Figure 4.2 showcases the first scenario, where the first two columns represent the training pairs for the DensePose-based ControlNet. The second scenario is of high interest to us, as it allows mapping from the SMPL-X avatar to the generated image that is well aligned with its control (see Figure 4.3).

To summarize, our ControlNet training and then conditional human generation process for synthetic dataset creation looks as follows:

- Predict DensePose part indices on the unlabeled human dataset utilizing accurate predictor.
- Train ControlNet adapter model for conditional human generation based on well-aligned DensePose modality.
- Given any SMPL or SMPL-X avatar, render it to the DensePose and provide it as input to the ControlNet to generate the human image accurately aligned with the avatar.

One might argue that when rendering an avatar to the DensePose part indices map, we might lose lots of 3D information about the human pose. Yet, we show that due to the constrained nature of the human body and the text and image prompting that we explain next, we can generate quite accurately aligned human images with the provided avatar label both in 2D and 3D.

4.1.3 Handling occlusions

FIGURE 4.4: Handling occlusions in the generated images: original DensePose prediction (top left), DensePose rendered from avatar prediction (bottom left), generation w/o occlusion awareness (bottom right), generation w/ occlusion awareness (top right) - right foot is properly occluded

Another aspect we should consider is the occlusion of the human body by the surrounding objects. The DensePose predictor [Güler, Neverova, and Kokkinos, 2018] learned to predict background in places where the human body is occluded, except for self-occlusions or occlusions by the human cloth. This property is transferred to the ControlNet, which generates images from the DensePose. On the other hand, avatars are always given in full length, if not limited by the image size. Thus, when rendered, DensePose produced from the avatar ignores any occlusion. All this results in the undesired effect that human bodies are always generated unobstructed by our pipeline because avatars are always rendered to full, unoccluded DensePose masks, and ControlNet learned that if the input control is full, the generation should be unoccluded as well.

To alleviate this issue, we transfer occlusions from the original DensePose estimate on the image to the DensePose obtained from the avatar. We keep DensePose rendered from the avatar only in the area under the dilated original DensePose estimate. This accounts for small misalignments in both masks and simultaneously

removes large chunks of rendered DensePose masks that should be occluded. The effect can be seen in the figure 4.4.

4.2 Label generation process

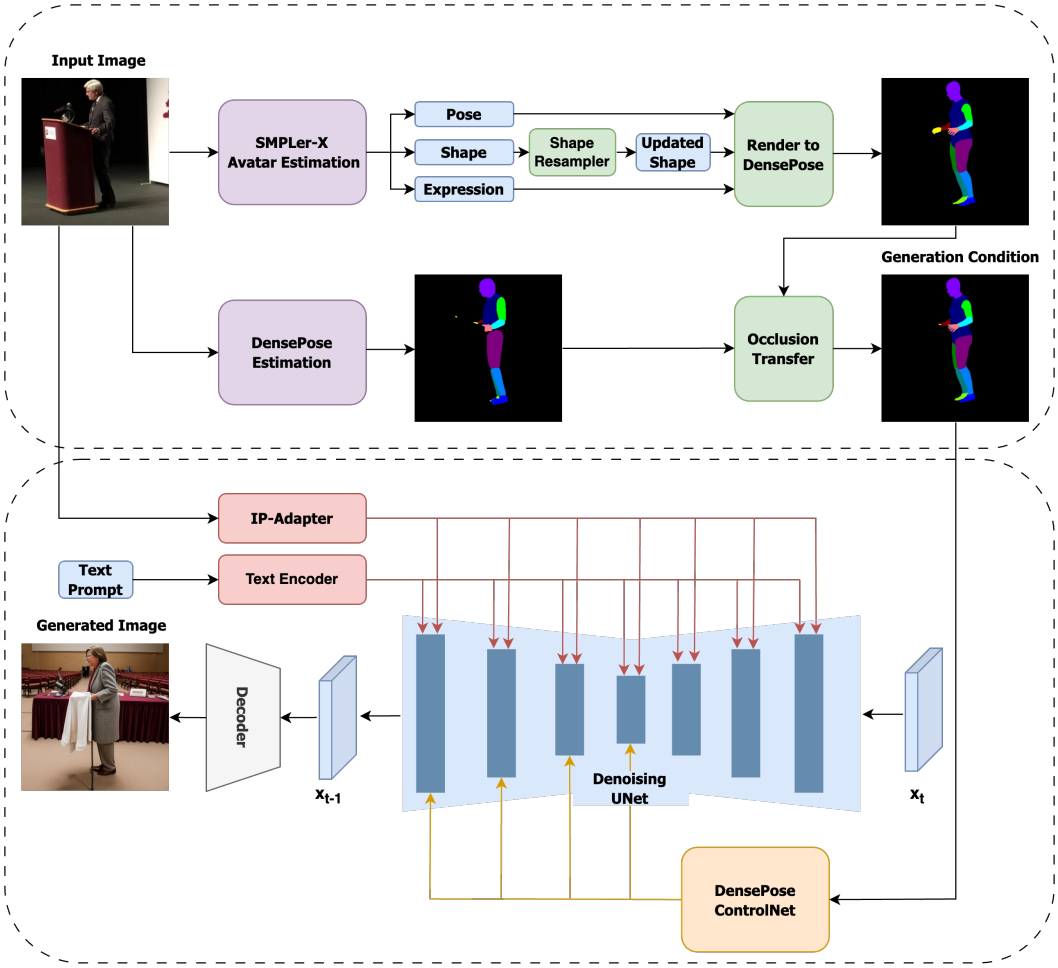

FIGURE 4.5: Conditional human generation pipeline that preserves the distribution of a base dataset. Top: We obtain the SMPL-X parameters from the state-of-the-art human mesh estimation method and transform them to the occluded DensePose representation. Bottom: Pose and shape condition are fused via ControlNet, while other aspects are encoded into the generation via cross-attention mechanism from the original image and text prompt

4.2.1 Obtaining real-world labels distribution

Here, we describe the label generation process. We chose to sample the parameters of the SMPL-X [Pavlakos et al., 2019] avatar from the SMPL [Loper et al., 2015] family of parametric human body models, as it conveniently captures all the information about the pose and shape of the human body, hands, and facial expression. Knowing the true avatar associated with the human in the image, we can extract multiple modalities to train the downstream model.

We need to establish a label sampling process for the SMPL-X avatars, such that the distribution of labels reflects the real-world distribution of the human poses, shapes, and expressions. One approach could be to model the parameters of SMPL-X avatars with simple distributions from which we can sample. However, there is no theoretical guarantee that any choice of distributions for pose or shape will reflect the real distribution of avatars. Moreover, if we use latent diffusion models like Stable Diffusion, we can control the generation process by the text prompt as a part of the condition. Establishing a reasonable text prompt for the sampled avatar is non-trivial and cannot be directly obtained for a random avatar.

To alleviate the abovementioned problem, we directly obtain SMPL-X avatars from real-world images with the existing predictor. Moreover, we select a dataset with text captions already available. Starting with the large dataset of human images and corresponding text captions LAION-Face [Zheng et al., 2022], which is a subset of larger LAION-400M [Schuhmann et al., 2021], we estimate the human mesh SMPL-X parameter using state-of-the-art method SMPLer-X [Cai et al., 2023]. Obtained avatars cannot be directly utilized as pseudo-ground truth labels for the given images because of their lack of accuracy. Yet, they reflect the real-world distribution of avatars, and if we can generate images that follow those avatars precisely, a desired synthetic dataset can be obtained.

This process can be seen as a dataset refinement. Firstly, we make predicting $\hat{y}_i$ on a large-scale unlabeled dataset $\{x_1, x_2, ..., x_m\}$ with the existing estimation method to obtain inaccurate image-label pairs with the predictions $\hat{D}$:

$$\hat{D} = \{(x_1, \hat{y}_1), (x_2, \hat{y}_2), ..., (x_m, \hat{y}_m)\},$$

and then use controllable image generation procedure $G(\hat{y}_i)$ to obtain images $x_i^{(s)}$ and compose from them synthetic dataset $D^{(s)}$:

$$D^{(s)} = \{(x_1^{(s)}, \hat{y}_1), (x_2^{(s)}, \hat{y}_2), ..., (x_m^{(s)}, \hat{y}_m)\}$$

Figure 4.5 shows the overview of the pipeline for the controllable synthetic human image $x_i^{(s)}$ generation with the preservation of pose, shape, and identity distribution from the base image x_i.

4.2.2 Changing the distribution

Obtaining labels from the model's predictions on the large-scale dataset can reflect the real-world label distribution but simultaneously capture biases learned by the model. For the SMPLer-X human mesh estimator, we identify such bias in the shape estimation, which is usually close to the mean shape, meaning that shape parameters $\vec{\beta}$ are close to 0. This bias in the model can be explained by how SMPLer-X is trained. A large portion of the training data consists of only 3D skeletons, from which estimating the true shape of the avatar is very hard.

To partially alleviate the problem, we change the shape parameters for each image pair by independently sampling shape coefficients. Since parameters $\vec{\beta}$ are coefficients near principal components, we argue that the true distribution of human shapes can be approximated by sampling each β_i component independently from a standard normal distribution and scaling it by the variance explained by the corresponding principal component. Sampling the shape conditional on the other factors, such as text prompt or pose, as well as alleviating other biases that might be present in the avatar predictor we leave as the future work.

4.2.3 IP-adapter conditioning

FIGURE 4.6: Introduction of the original image as an image prompt through IP-Adapter. Left to right: original image, DensePose of predicted avatar, generated image w/o image prompt, generated image w/ image prompt. IP-adapter allows to capture better such concepts as clothing, gender, background, etc.

Additionally, we want to preserve the distribution of images as close as possible to the original large-scale dataset LAION-Faces and remove biases that might be introduced with the diffusion model, e.g., human skin color, lighting conditions, background, and clothing. Part of the information can be encoded in the text prompt, but realistically, not all aspects of the image can be captured through a textual description.

Luckily, each generated image $x_i^{(s)}$ is controlled by the label $\hat{y}_i$, which in turn was estimated from the original image x_i. Thus, we introduce additional conditioning in the form of original image x_i into the diffusion model through the IP-Adapter [Ye et al., 2023]. IP-Adapter uses highly compressed CLIP-based [Radford et al., 2021] representation of input image, which carries semantic information but loses low-level details. This allows us to achieve conditioning on the human pose and shape from our label while transferring semantics from the original image x_i to the generated $x_i^{(s)}$.

4.3 Avatar estimation

To showcase the applicability of our synthetic dataset, we utilize it for the downstream task of human mesh estimation in the form of SMPL-X parameter prediction. To fairly compare against the competitors, we choose the architecture of the state-of-the-art model SMPLer-X [Cai et al., 2023] and train it on our dataset.

4.3.1 Virtual Markers Prediction

We propose a minor architectural change to the SMPLer-X architecture to help shape predictions. Originally, SMPLer-X first predicted the location of 3D joints, which were then used as input to the pose parameters prediction head. The joints are directly supervised in addition to the avatar parameters to improve the stability and speed of training.

Potentially, using synthetic data and precise knowledge of ground truth avatar, one could build an architecture where each individual vertex is predicted in a similar manner to joints. However, this might be redundant and very inefficient as there is a large number of vertices ($\sim$10k) compared to tens of joints, and a 3D heatmap needs to be predicted for each vertex.

Therefore, following VirtualMarker [Ma et al., 2023], we introduce additional intermediate predictions into the model in the form of markers. Markers are simply a small set $M = 64$ of vertices fixed on the mesh surface. Compared to joints, they are influenced more by the shape variations of the human body, thus helping with predicting shape parameters. Yet, they are sparsely located on the mesh surface, making them an efficient representation of the human body. We predict joints and markers at the same time and send markers to the shape parameters prediction head as additional input. The loss functions on markers are analogous to the joint loss functions described in chapter 3. Markers are supervised both in 3D and 2D, directly and after reconstructing them from predicted avatars.

$$\begin{aligned} \mathcal{L}_{3D}^{m} &= \sum_{i} ||m_i^{3D} - \hat{m}_i^{3D}||_1 \\ \mathcal{L}_{2D}^{m} &= \sum_{i} ||m_i^{2D} - \hat{m}_i^{2D}||_1, \end{aligned} \tag{4.1}$$

Chapter 5

Experiments and Results

Below, we provide a list of experiments and results that showcase the effectiveness of using synthetic data for human mesh estimation. We start by describing the setup and specific parameters of various pipeline components, like ContolNet training, controllable generation, and finally, SMPLer-X model training. Then, we move to the exploration datasets and metrics used for evaluation and compare the results of various approaches and proposed improvements.

5.1 Experimental Setup

The pipeline consists of multiple stages, and here, we provide the necessary experimental setups for each of them. All the experiments are implemented in PyTorch [Paszke et al., 2019] framework, and the Stable Diffusion Model inference code is called through the Diffusers Library [Platen et al., 2022].

Also, we use a subset of the LAION-Faces dataset [Zheng et al., 2022] throughout our experiments. We filter images where humans are visible in large enough resolution, and at least half of their body is visible based on the detected 2D joints and arrive at $\sim$300k images. The dataset is used both for ControlNets training and as a basis dataset for deriving synthetic datasets.

5.1.1 ControlNet learning

We train the ControlNet [Zhang, Rao, and Agrawala, 2023] model for the Stable Diffusion 1.5 base model on two different modalities - color-coded SMPL-X avatars and DensePose part index segmentation. We train on the subset of the LAION-Faces dataset labeled with SMPLer-X (ViT-L) for the color-coded avatar representation. For DensePose, we use the original model [Güler, Neverova, and Kokkinos, 2018] based on the Mask R-CNN architecture [He et al., 2018].

The training was done on 4 Nvidia RTX 3090 GPUs with an effective cumulative batch size of 32 for 2 training epochs. ControlNet is known for its sudden convergence effect, which was observed after the end of the first epoch. For training, the images were augmented by random scaling, cropping, and rotations and resized to the size of 512px by 512px.

5.1.2 Dataset generation

The ControlNet models are used to allow for controllable generation with the SD model. Once again, we predict avatars on the LAION-Faces dataset and render the predicted avatars to the color-coded and DensePose part index representation. The controllable generation pipeline is used for image generation. We use the DDIM

scheduler [Song, Meng, and Ermon, 2022] and set the generation length to 20 iterations. Also, when used, the IP-Adapter influence scale is set to 0.8. We also employ the FreeU [Si et al., 2023] method for improving the generation quality and DeepCache [Ma, Fang, and Wang, 2023] for speeding up the generation by caching intermediate UNet activations between different denoising steps. Both of the approaches do not require any base model training and can be easily integrated with other components used for conditioning. The generation image size of the smaller image side is set to 512, while the larger one is set such that the aspect ratio of the original image is preserved.

We call the dataset generated with the color-coded SMPL-X avatar and DensePose based ControlNets LAION-Synth-300k-CC and LAION-Synth-300k-DP, respectively.

5.1.3 SMPLer-X training

We train the SMPLer-X model only on our synthetically generated dataset, which removes a lot of the complexity of the original SMPLer-X training. Most of the experiments are done on the model with ViT-S backbone to compare it fairly against individual dataset training setups of the original SMPLer-X model. We also train the model with ViT-H backbone to compare against state-of-the-art SMPLer-X, yet results on the individual real-world datasets are not available. The backbones are initialized with the weights of the ViTPose model [Xu et al., 2022], which has learned to predict 2D human joints in the image and is the closest related task to human mesh estimation with such large-scale foundational models available.

In all setups, the model was trained on 6 Nvidia RTX 3090 GPUs for 10 epochs, with a training cumulative batch size of 16 and an input image shape of 512 by 384 pixels cropped around the human bounding box. We use Adam optimizer [Kingma and Ba, 2017] for model training.

5.2 Results

5.2.1 Evaluation datasets

Following the SMPLer-X work, we evaluate our method on 3 different real-world datasets with ground truth SMPL or SMPL-X avatars available: EHF [Pavlakos et al., 2019], EgoBody (EgoSet) [Zhang et al., 2022], UBody [Lin et al., 2023].

EHF

The dataset consists of 100 images of a single subject in the laboratory environment with ground truth avatar parameters available. The poses are diverse, but the shape is not. The dataset has only a test subset and no training or validation subsets.

EgoBody (EgoSet)

EgoBody is a large-scale dataset with the subject's videos captured from the first-person view. For efficient evaluation and to skip consequent frames that are almost identical, we subsample only 1 in 100 frames from the test set, arriving at 3000 test images.

UBody

Similarly to EgoBody, UBody is a large-scale dataset with subject videos that covers 15 real-world scenarios. Again, for efficiency reasons, we subsample only 1 in 1000 frames from the test set, arriving at 9000 test images.

SSP3D

Additionally, we do ablations on the SSP3D [Sengupta, Budvytis, and Cipolla, 2020] dataset, which includes 311 images of athletes images. The main value of the dataset is in its diverse poses and shapes that athletes possess. Thus, it is great at evaluating the shape estimation capabilities of the models.

The SSP3D dataset was included in the training dataset of the original SMPLer-X model. Therefore, we cannot use the metrics of the original SMPLer-X on this dataset to compare it to our model. We only use SSP3D for internal ablations.

5.2.2 Evaluation metrics

We follow the standard practice of evaluating human mesh estimation methods. Comparison in the parameters space of SMPL-X avatar is meaningless, thus we evaluate the differences of avatars build from those parameters. Two primary metrics include Per-Vertex Error (PVE) - the mean difference between corresponding vertices of true and predicted avatar aligned by the location of root joint, and Procrustes-Aligned Per-Vertex Error (PA-PVE) - which is simply PVE after finding optimal rigid alignment between two avatars.

$$\begin{aligned} \text{PVE} &= \frac{1}{|V|}\sum_{i}^{|V|} ||\vec{v}_{gt} - \vec{v}_{pred}|| \\ \text{PA-PVE} &= \frac{1}{|V|}\sum_{i}^{|V|} ||\vec{v}_{gt}^{(PA)} - \vec{v}_{pred}^{(PA)}|| \end{aligned} \tag{5.1}$$

For the SSP3D dataset, we only want to measure the accuracy of the shape. For that purpose, we calculate PVE on the avatars in zero pose, meaning that we set pose parameters to zero and keep only the shape parameters untouched when constructing the avatars. This effectively takes into account only shape during evaluation. We call this metric PVE-S, standing for Per-Vertex Error Shape.

5.2.3 Experiments results

We start with the comparison of the conditioning modality for the ControlNet, namely comparing two versions of synthetic datasets, LAION-Synth-300k-CC and LAION-Synth-300k-DP. For this purpose, we train on both datasets the SMPLer-X model with ViT-S backbone and compare on three proposed evaluation datasets. The results can be seen in table 5.1. The DensePose-based dataset clearly outperforms the Color-Coded Avatar-based dataset, verifying visual clues that poor alignment during ControlNet training results in poor controllability. The next experiments will involve only DensePose-based conditioning.

TABLE 5.1: The quantitative evaluation of the quality of synthetic data generated by different conditioning modalities. LAION-Synth-300k-CC represents the dataset created by ControlNet, taking as input color-coded SMPL-X avatars, while LAION-Synth-300k-DP takes as input DensePose estimates. Datasets are compared by reporting metrics of the SMPLer-X (ViT-S) model trained on them

Synthetic Dataset ↓	EHF		EgoBody		UBody	
	PVE	PA-PVE	PVE	PA-PVE	PVE	PA-PVE
LAION-Synth-300k-CC	120.51	93.12	168.32	112.30	154.29	97.16
LAION-Synth-300k-DP	97.84	60.94	129.81	81.62	112.78	54.65

Next, we evaluate the changes aimed at fixing the label distribution and infusing additional training signals into the model. In particular, we test the impact of resampling the shape parameters and generating the LAION-Synth-300k-DP-Shape-Resampled version of the dataset and the introduction of virtual markers into the training signal, as well as an additional input to the shape parameters prediction head. The results can be seen in table 5.2. Adding markers to the training and resampling shape distribution improves the quality on multiple evaluation dataset datasets, but decreases metrics on the EHF. This can be explained by a single subject present in this dataset, with a shape close to the mean human shape. Decreases in errors on the SSP3D dataset, where shape variations in the human body are the most prominent, prove the effectiveness of proposed improvements for the shape estimation capabilities of the model.

TABLE 5.2: The quantitative evaluation of fixing shape distribution in the synthetic dataset and introduction of additional virtual markers modality. Reported metrics represent the quality of the SMPLer-X (ViT-S) model trained with the specified method. S. r. stands for resampled shape in dataset distribution

Method ↓	EHF		EgoBody		UBody		SSP3D		
	PVE	PA-PVE	PVE	PA-PVE	PVE	PA-PVE	PVE	PA-PVE	PVE-S
Baseline	97.84	60.94	129.81	81.62	112.78	54.65	116.26	76.39	53.58
w/ s.r.	102.37	60.55	129.30	81.63	112.33	54.62	113.82	74.92	50.15
w/ markers	103.15	60.67	127.23	79.70	112.42	53.15	111.27	73.63	49.82
w/ s.r. + markers	105.59	60.63	127.01	79.46	111.73	52.70	106.85	72.07	50.9

Moreover, we compare our model (without markers) with the metrics of models trained on individual synthetic datasets. For comparison, we choose five synthetic datasets created in the gaming engines BEDLAM [Black et al., 2023b], SynBody [Yang et al., 2023], AGORA [Patel et al., 2021b], GTA-Human II [Cai et al., 2022], and SPEC [Kocabas et al., 2022]. The sizes of datasets range from 72k instances in SPEC to 1802k instances in GTA-Human II. For all except our dataset, the metrics are taken from [Cai et al., 2023] and are reported in table 5.3. Our dataset shows comparable and sometimes even better performance compared to other synthetic data, which shows its effectiveness in training human mesh estimation models.

Finally, we train a SMPLer-X model with ViT-H backbone to compare against the state-of-the-art SMPLer-X model. Importantly, the original SMPLer-X is trained on 32 datasets and learns on training subsets of EgoBody and UBody as well. This makes the comparison of validation subsets of these datasets unfair with regard to our model, which never saw any of these datasets. Yet, we present the metrics in

TABLE 5.3: The quantitative evaluation of model trained on our LAION-Synth-300k-DP-Shape-Resampled dataset against other synthetic datasets.

Training Dataset ↓	EHF PVE	EgoBody PVE	UBody PVE
BEDLAM	<u>81.10</u>	<u>109.10</u>	132.50
SynBody	112.90	136.60	144.60
AGORA	164.60	138.40	128.40
GTA-Human II	126.00	139.20	143.70
SPEC	197.80	154.80	146.10
Synth-300k-DP-SR	102.37	129.30	<u>112.33</u>

table 5.4 and show the gap that still exists between our best model and state-of-the-art. We do not compare on the SSP3D dataset as it was part of the training data for the original SMPLer-X. Qualitative results of our best model predictions on the unseen images are shown in figure 5.1.

TABLE 5.4: The quantitative evaluation of model trained on our LAION-Synth-300k-DP-Shape-Resampled dataset with ViT-H backbone against state-of-the-art SMPLer-X model trained on 32 datasets.

Training Dataset ↓	EHF PVE	EgoBody PVE	UBody PVE
32 datasets	<u>56.80</u>	<u>59.50</u>	<u>54.50</u>
Synth-300k-DP-SR	74.31	99.12	90.49

FIGURE 5.1: The qualitative results of SMPLer-X ViT-H model trained on Synth-300k-DP-SR dataset

Chapter 6

Conclusions

6.1 Contribution

In this work, we propose a novel approach to generating a dataset for human pose and shape estimation tasks in the form of SMPL-X avatar parameters prediction. To our knowledge, we are the first to utilize purely synthetic data generated by the Latent Diffusion Model through a conditional generation process for this task.

- We develop a method for controllable synthetic human generation based on the DensePose representation and show its superiority compared to naive generation based on avatar renders.
- We train the human mesh estimation model SMPLer-X purely on our synthetic data. We show that the results of a model trained on our dataset are comparable and sometimes even superior to the model trained on other synthetic datasets utilized for human avatar regression methods.
- In contrast with other synthetic datasets, our dataset generation process does not involve a human for its creation. Also, it is not limited in diversity by the attributes predefined by the computer graphics artist.

6.2 Limitations & Future Work

Due to limited computational resources, we did not experiment extensively with the scaling laws for our dataset size. Also, we utilized Stable Diffusion 1.5, which might not be optimal considering the existence of a larger model SDXL [Podell et al., 2023], which requires a much more computational budget for ControlNet training and further generation but consistently provides better results. Moreover, we did not propose any human identity preservation in our dataset creation process, as it might leak into the generated image, especially with the usage of the original image as a prompt for the generation.

In future work, we plan to mitigate those limitations. We plan to extend the model to generate and then perform human mesh estimation on multiple people in the image at the same time. Also, we plan on extending the controllability of generation for camera parameters as well, which might help in the camera estimation process as a subtask of human mesh recovery.

Lastly, we plan to resample other factors in the generation process, similar to what we did with shape parameters. We plan to resample pose distribution for complex poses and evaluate the method on datasets with non-trivial human pose distributions. Also, more careful evaluation of human hands and head estimation capabilities is needed.

Bibliography

Azizi, Shekoofeh et al. (2023). *Synthetic Data from Diffusion Models Improves ImageNet Classification*. arXiv: 2304.08466 [cs.CV].

Baradel, Fabien et al. (2024). "Multi-HMR: Multi-Person Whole-Body Human Mesh Recovery in a Single Shot". In: *arXiv*.

Black, Michael J. et al. (June 2023a). "BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion". In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737.

— (2023b). *BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion*. arXiv: 2306.16940 [cs.CV].

Cai, Zhongang et al. (2022). *Playing for 3D Human Recovery*. arXiv: 2110.07588 [cs.CV].

Cai, Zhongang et al. (2023). *SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation*. arXiv: 2309.17448 [cs.CV].

Chang, Di et al. (2023). "MagicDance: Realistic Human Dance Video Generation with Motions & Facial Expressions Transfer". In: *arXiv preprint arXiv:2311.12052*.

Dhariwal, Prafulla and Alex Nichol (2021). *Diffusion Models Beat GANs on Image Synthesis*. arXiv: 2105.05233 [cs.LG].

Dosovitskiy, Alexey et al. (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV].

Erfanian Ebadi, Salehe et al. (2022). "PSP-HDRI+: A Synthetic Dataset Generator for Pre-Training of Human-Centric Computer Vision Models". In: *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*.

Fu, Jianglin et al. (2022). *StyleGAN-Human: A Data-Centric Odyssey of Human Generation*. arXiv: 2204.11823 [cs.CV].

Goodfellow, Ian J. et al. (2014). *Generative Adversarial Networks*. arXiv: 1406.2661 [stat.ML].

Güler, Rıza Alp, Natalia Neverova, and Iasonas Kokkinos (2018). *DensePose: Dense Human Pose Estimation In The Wild*. arXiv: 1802.00434 [cs.CV].

He, Kaiming et al. (2018). *Mask R-CNN*. arXiv: 1703.06870 [cs.CV].

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). *Denoising Diffusion Probabilistic Models*. arXiv: 2006.11239 [cs.LG].

Ho, Jonathan and Tim Salimans (2022). *Classifier-Free Diffusion Guidance*. arXiv: 2207.12598 [cs.LG].

Hu, Edward J. et al. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv: 2106.09685 [cs.CL].

Huang, Zhen et al. (2023). *FIRST: A Million-Entry Dataset for Text-Driven Fashion Synthesis and Design*. arXiv: 2311.07414 [cs.CV].

Ionescu, Catalin et al. (2014). "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, pp. 1325–1339.

Karras, Tero et al. (2020). *Analyzing and Improving the Image Quality of StyleGAN*. arXiv: 1912.04958 [cs.CV].

Kingma, Diederik P. and Jimmy Ba (2017). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Kingma, Diederik P. and Max Welling (2019). "An Introduction to Variational Autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4, 307–392. ISSN: 1935-8245. DOI: 10.1561/2200000056. URL: http://dx.doi.org/10.1561/2200000056.

Kocabas, Muhammed et al. (2022). *SPEC: Seeing People in the Wild with an Estimated Camera*. arXiv: 2110.00620 [cs.CV].

Lin, Jing et al. (2023). *One-Stage 3D Whole-Body Mesh Recovery with Component Aware Transformer*. arXiv: 2303.16160 [cs.CV].

Lin, Tsung-Yi et al. (2015). *Microsoft COCO: Common Objects in Context*. arXiv: 1405.0312 [cs.CV].

Liu, Lijuan et al. (2023). "Towards Garment Sewing Pattern Reconstruction from a Single Image". In: *ACM Transactions on Graphics (SIGGRAPH Asia)*.

Loper, Matthew et al. (Oct. 2015). "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6, 248:1–248:16. DOI: 10.1145/2816795.2818013.

Ma, Qianli et al. (2020). "Learning to Dress 3D People in Generative Clothing". In: *Computer Vision and Pattern Recognition (CVPR)*.

Ma, Xiaoxuan et al. (2023). *3D Human Mesh Estimation from Virtual Markers*. arXiv: 2303.11726 [cs.CV].

Ma, Xinyin, Gongfan Fang, and Xinchao Wang (2023). *DeepCache: Accelerating Diffusion Models for Free*. arXiv: 2312.00858 [cs.CV].

Marcard, Timo von et al. (2018). "Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera". In: *European Conference on Computer Vision (ECCV)*.

Mehta, Dushyant et al. (2017). "Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision". In: *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE. URL: http://gvv.mpi-inf.mpg.de/3dhp_dataset.

Moon, Gyeongsik, Hongsuk Choi, and Kyoung Mu Lee (2022). "NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets". In: *Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Moon, Gyeongsik et al. (2023). "Three Recipes for Better 3D Pseudo-GTs of 3D Human Mesh Estimation in the Wild". In: *Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Mou, Chong et al. (2023). *T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models*. arXiv: 2302.08453 [cs.CV].

Paszke, Adam et al. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv: 1912.01703 [cs.LG].

Patel, Priyanka et al. (June 2021a). "AGORA: Avatars in Geography Optimized for Regression Analysis". In: *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

— (2021b). *AGORA: Avatars in Geography Optimized for Regression Analysis*. arXiv: 2104.14643 [cs.CV].

Pavlakos, Georgios et al. (2019). "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985.

Platen, Patrick von et al. (2022). *Diffusers: State-of-the-art diffusion models*. https://github.com/huggingface/diffusers.

Podell, Dustin et al. (2023). *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. arXiv: 2307.01952 [cs.CV].

Radford, Alec et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. arXiv: 2103.00020 [cs.CV].

Rombach, Robin et al. (2022). *High-Resolution Image Synthesis with Latent Diffusion Models*. arXiv: 2112.10752 [cs.CV].

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv: 1505.04597 [cs.CV].

Sariyildiz, Mert Bulent et al. (2023). *Fake it till you make it: Learning transferable representations from synthetic ImageNet clones*. arXiv: 2212.08420 [cs.CV].

Schuhmann, Christoph et al. (2021). *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*. arXiv: 2111.02114 [cs.CV].

Sengupta, Akash, Ignas Budvytis, and Roberto Cipolla (2020). *Synthetic Training for Accurate 3D Human Pose and Shape Estimation in the Wild*. arXiv: 2009.10013 [cs.CV].

SG_161222 (2023). *Realistic Vision v5.1*. URL: https://civitai.com/models/4201?modelVersionId=130072.

Shipard, Jordan et al. (2023). *Diversity is Definitely Needed: Improving Model-Agnostic Zero-shot Classification via Stable Diffusion*. arXiv: 2302.03298 [cs.CV].

Shlens, Jonathon (2014). *A Tutorial on Principal Component Analysis*. arXiv: 1404.1100 [cs.LG].

Si, Chenyang et al. (2023). *FreeU: Free Lunch in Diffusion U-Net*. arXiv: 2309.11497 [cs.CV].

Song, Jiaming, Chenlin Meng, and Stefano Ermon (2022). *Denoising Diffusion Implicit Models*. arXiv: 2010.02502 [cs.LG].

Tian, Yonglong et al. (2023). *StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners*. arXiv: 2306.00984 [cs.CV].

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL].

Xu, Yufei et al. (2022). *ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation*. arXiv: 2204.12484 [cs.CV].

Yan, Haonan et al. (2021). "UltraPose: Synthesizing Dense Pose With 1 Billion Points by Human-Body Decoupling 3D Model". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10891–10900.

Yang, Zhitao et al. (2023). *SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling*. arXiv: 2303.17368 [cs.CV].

Ye, Hu et al. (2023). *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. arXiv: 2308.06721 [cs.CV].

Zhang, Hongwen et al. (2021). *PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop*. arXiv: 2103.16507 [cs.CV].

Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala (2023). *Adding Conditional Control to Text-to-Image Diffusion Models*. arXiv: 2302.05543 [cs.CV].

Zhang, Siwei et al. (2022). *EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices*. arXiv: 2112.07642 [cs.CV].

Zheng, Yinglin et al. (2022). "General facial representation learning in a visual-linguistic manner". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18697–18709.