Towards Robust and Expressive Human Modelling

Pang Hui En

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2026

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

07/08/2025

Date

Pang Hui En

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

07/08/2025

Date

Prof. Zhang Tianwei

Authorship Attribution Statement

This thesis contains material from three papers published in the peer-reviewed conferences, in which I am listed as an author.

Chapter 3 is published as **Hui En Pang**, Zhongang Cai, Lei Yang, Tianwei Zhang, Ziwei Liu. Benchmarking and Analyzing 3D Human Pose and Shape Estimation Beyond Algorithms. In Proceedings of Neural Information Processing Systems (NeurIPS Dataset and Benchmark Track), 2022.

The contributions of the co-authors are as follows:

- I served as the lead author, conducted experiments and analysis, and prepared the manuscript.
- Dr. Cai contributed to the method design and project conception.
- Dr. Cai and Dr. Yang provided technical guidance and key suggestions to improve the manuscript.
- Prof. Zhang Tianwei and Prof. Liu Ziwei shaped the initial project direction and offered invaluable insights.
- We all participated in discussions and contributed to manuscript revision.

Chapter 4 is published as **Hui En Pang**, Zhongang Cai, Lei Yang, Tianwei Zhang, Qingyi Tao, Zhonghua Wu, Ziwei Liu. Towards Robust and Expressive Whole-body Human Pose and Shape Estimation. In Proceedings of Neural Information Processing Systems, 2023.

The contributions of the co-authors are as follows:

- I served as the lead author, conducted experiments and analysis, and prepared the manuscript.
- Dr. Cai and Dr. Yang provided technical guidance and key suggestions to improve the manuscript.
- Prof. Zhang Tianwei and Prof. Liu Ziwei shaped the initial project direction and offered invaluable insights.
- We all participated in discussions and contributed to manuscript revision.

Chapter 5 is published as **Hui En Pang**, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, Ziwei Liu. Disco4D: Disentangled 4D Human Generation and Animation from a Single Image. In Conference on Computer Vision and Pattern Recognition, 2025.

The contributions of the co-authors are as follows:

- I served as the lead author, conducted experiments and analysis, and prepared the manuscript.

- Dr. Cai and Dr. Yang provided technical guidance and key suggestions to improve the manuscript.
- Prof. Zhang Tianwei and Prof. Liu Ziwei shaped the initial project direction and offered invaluable insights.
- We all participated in discussions and contributed to manuscript revision.

07/08/2025

.....

Date

.....

Pang Hui En

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Zhang Tianwei, for his patient guidance, encouragement, and invaluable advice throughout my PhD journey. I am thankful for his consistent mentorship and regular check-ins, and I have been inspired by his strong commitment to the growth of his students. His support during challenging periods and his insightful feedback at every stage of my PhD have been instrumental in helping me mature as a researcher.

In addition, I would like to extend my sincere thanks to Prof. Liu Ziwei for his collaboration and for playing a crucial role in many of my works. His mentorship and constructive feedback have greatly shaped the direction and quality of my research.

I am grateful to the SenseTime-NTU Talent Programme, without which I would not have had the opportunity to pursue this PhD. My two years at SenseTime were especially meaningful, as I had the privilege of working under the guidance of Dr. Cai Zhongang and Dr. Yang Lei. Their expert advice and encouragement were pivotal to many of my works, deepening my interest in human modelling and broadening my perspective on impactful research. I also received valuable support from Liu Shuai, Dr. Wong Yuen Fei, Dr. Tao Qingyi, Dr. Wu Zhonghua, and my fellow IPP students during my time there. I would like to thank SenseTime for both their financial support and the computational resources that were essential for conducting my experiments.

I am also grateful to AI Singapore for awarding me the AISG PhD Fellowship, and to Prof. Low Bryan for his advice and guidance. I sincerely thank AI Singapore for their financial support that made my research possible.

Lastly, I want to express my heartfelt appreciation to my family for their unwavering love and support. Thank you to my parents and brother for always encouraging me to pursue my interests to the fullest, and to my partner and close friends who have been by my side since the start of this journey. This PhD has been full of ups

and downs and moments of uncertainty, and I could not have completed it without their constant encouragement and belief in me.

"How lucky I am to have something that makes saying goodbye so hard."

—Winnie the Pooh

To my family

Abstract

This thesis advances robust and expressive 3D human modeling from a single image, addressing two core challenges: (1) recovering accurate human pose and shape beneath clothing, and (2) generating animatable, disentangled 3D avatars with reusable assets. It is structured around four works that progressively tackle estimation, representation, and animation of clothed humans.

The first work presents a comprehensive benchmarking study of human mesh recovery (HMR). While prior research focuses primarily on algorithm design, this study systematically examines the impact of datasets, backbone architectures, and training strategies. It reveals that data diversity, task-aligned architectures, and well-designed training protocols play a critical role in generalization, and establishes strong baselines and practical guidelines for fair evaluation.

The second work, RoboSMPL-X, addresses whole-body pose and shape estimation from a single image, including body, hands, and face. We identify bounding-box quality as a key bottleneck in real-world settings, and propose a framework that improves spatial localization, learns augmentation-invariant features through contrastive learning, and enforces image-space pixel alignment of the predicted mesh, resulting in improved robustness under challenging conditions.

The third work, Disco4D, shifts from estimation to generative modeling. It introduces a layered representation that combines SMPL-X body modeling with Gaussian Splatting, enabling separate modeling of body and clothing. This formulation supports temporally coherent 4D animation and facilitates asset-level manipulation, allowing clothing to be extracted, reused, and edited across different identities.

The fourth work, ReposeHuman, focuses on generating customizable and animatable avatars with explicit disentanglement of body and clothing. To overcome the limitations of single-image supervision, it leverages video diffusion models to synthesize canonical multi-view observations, which provide strong geometric supervision for

reconstructing disentangled 3D human meshes. A large-scale dataset of animatable, disentangled human meshes is introduced to support this task.

Together, these works form a cohesive pipeline from robust estimation to expressive generation. By addressing both robustness and disentangled representation, this thesis contributes scalable methods for creating high-fidelity, controllable, and animatable digital humans from minimal input.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

3D human modeling and generation has garnered significant attention due to its wide range of applications in robotics, computer graphics, AR/VR, and more. Current approaches can be broadly divided into two categories: (1) human pose and shape estimation, which focuses on modeling the body beneath clothing using parametric models such as SMPL [1], and (2) clothed human digitization, which models both body and clothing to recover a complete dressed human mesh with fine-level details. For human pose and shape estimation (HPSE), recent works typically regress the parameters of a body model from monocular RGB images [2–5] or videos [6–8], leading to the term "human mesh recovery"[1]. A wide range of algorithms have been proposed to improve mesh recovery accuracy [9–17]. For clothed human digitization, a multitude of research efforts have been made on reconstructing 3D clothed human models from a single image [18–24, 24–27].

Taken together, these two categories address complementary facets of a broader question: *how can we obtain robust, expressive, and animatable 3D humans from in-the-wild visual input?* This thesis is organised around that question. Rather than treating the four technical contributions in isolation, the thesis positions them as successive stages of a broader pipeline, illustrated in Fig. 1.1: (1) understanding the factors that drive mesh-recovery performance beyond algorithms, (2) improving

[1]The terms "3D human pose and shape estimation" and "human mesh recovery" are used interchangeably in this thesis.

whole-body pose and shape estimation through robust and pixel-aligned recovery, (3) lifting a single image to a clothed and disentangled avatar, and (4) generating higher-fidelity canonical multi-view supervision for more controllable and animatable avatar reconstruction. Each subsequent chapter revisits this broader pipeline and identifies the bottleneck it addresses.

1.2 Challenges and Motivation

For existing HPSE methods, most efforts have centered on algorithmic improvements, with limited systematic investigation into other critical factors such as dataset design, network backbones, and training strategies. Moreover, many methods are tailored to standard benchmarks and often lack robustness in real-world conditions.

In contrast, clothed human digitization presents its own set of challenges. To support animation, the avatar must be reconstructed in a canonical pose, which is especially difficult from a single image due to the need to infer occluded or unseen regions, such as the back view, in monocular settings. Furthermore, most existing methods model the body and clothing jointly, which limits the animation, customization, and editability of individual components, all of which are crucial for building flexible and expressive avatars.

This section identifies the specific challenges addressed in this thesis and briefly motivates each corresponding study.

1.2.1 Lack of Systematic Benchmarking

Despite the progress in HPSE algorithms, there has been a lack of systematic investigation into other crucial factors that impact model performance. This includes the selection of datasets and their contributions, the choice of pretrained backbones for learning the HPSE model, and the impact of training strategies such as data augmentation and loss design. The influence of these factors on model performance and the optimal training configurations for achieving good HPSE models remain unclear.

This lack of understanding hampers the development of HPSE research in two main ways. Firstly, researchers may build and evaluate new algorithms using suboptimal training configurations, which fails to fully showcase the benefits of their inventions. Secondly, previous works often compare different algorithms or methods with varying training configurations, leading to unfair evaluations. This motivates our first study.

1.2.2 Robustness in Whole-Body Pose and Shape Estimation

Building on insights from our first study, where we identified optimal choices for datasets, network backbones, and training strategies for HPSE models, we shift our focus to algorithmic developments for improving robustness.

While whole-body human pose and shape estimation (HPSE) aims to jointly reconstruct the body, hands, and face, existing methods often struggle under in-the-wild conditions. Through a systematic benchmark of existing whole-body HPSE methods under various input augmentations, we observe that models are surprisingly sensitive to minor perturbations in scale and translation. Performance degrades significantly due to misaligned crops or inaccurate bounding boxes, which introduce errors that propagate along the kinematic chain.

The lack of robustness in existing whole-body pose and shape estimation methods highlights three critical aspects that can be improved upon: 1) accurate localization of the subject and its parts, 2) accurate extraction of useful features, and 3) accurate pixel alignment of outputs. To address these challenges, our second study introduces a robust and modular framework for whole-body HPSE.

1.2.3 Lack of Disentangled 3D Human Generation

Having established strong training foundations and improved robustness in human pose and shape estimation, we next turn to human digitization, where both body and clothing are modeled explicitly.

However, most existing 3D human generation pipelines represent the body and clothing as a single, entangled mesh. This design severely limits flexibility in animation, editing, and asset reuse. Without separable representations for individual

components—such as hair, shirts, pants, and shoes—it becomes challenging to enable layered rendering, simulate dynamic clothing behavior, or exercise fine-grained control over fashion items. These limitations hinder the creation of expressive and animatable digital humans. This motivates our third study, which introduces a disentangled representation for clothed human generation and animation.

1.2.4 Lack of Disentangled Datasets and Canonical Multi-View Supervision

While our previous work established a layered representation for disentangled body and clothing modeling, generating accurate canonical 3D avatars from single images remains a significant challenge, especially when the input depicts a non-standard or casual pose. General-purpose diffusion models are insufficient, as they often lack geometric grounding and fail to produce coherent views across different poses and angles. To achieve high-quality avatars with disentangled assets, we need multi-view images in canonical poses well-aligned with SMPL-X predictions.

Existing methods often rely on self-rotating video captures in canonical poses (e.g., T- or A-pose) in controlled environments to create animatable avatars, making it impractical for large-scale or casual use. Single-image diffusion-based animation methods [28–31] offer pose control but suffer from view inconsistencies, motion artifacts, and poor detail preservation, making them unsuitable for asset disentanglement. These issues arise from weak 2D training signals (e.g., keypoints) and limited supervision. Even dense-signal approaches like CHAMP [32] face misalignment and poor multi-view consistency, especially for occluded regions. Recent works such as CharacterGen [33] and EN3D [34] attempt canonical view synthesis but offer sparse views, weak control, and limited fidelity.

This motivates our fourth study: generating dense, multi-view canonical images from a single image using video diffusion models, enabling better alignment, disentanglement, and high-fidelity avatar reconstruction.

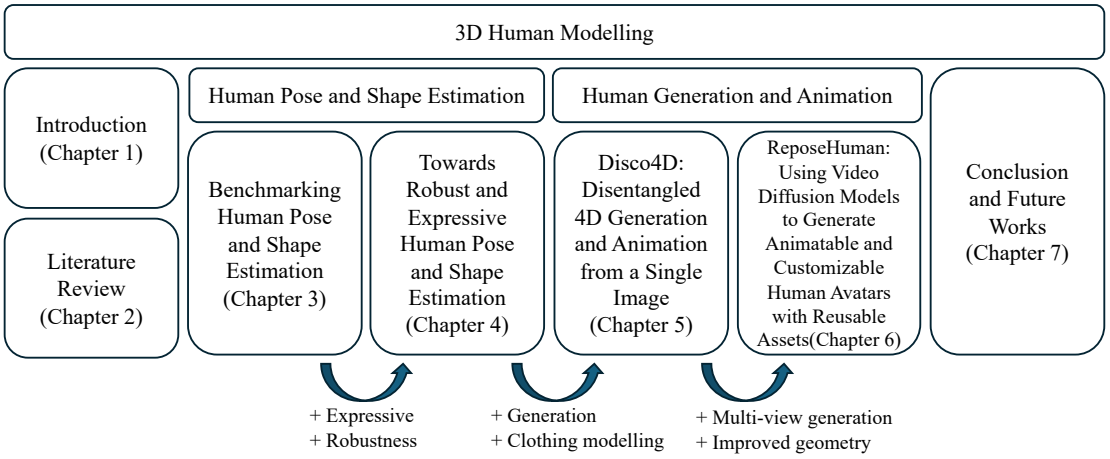


FIGURE 1.1: **Overview of Thesis.** We begin the thesis with an introduction to key research topics, challenges, and solutions. The technical sections introduce four works that support this thesis, which can be divided into two topics: human pose and shape estimation, and disentangled human generation and animation. Lastly, we conclude the thesis with a discussion of our proposed works and future directions.

1.3 Approaches

The overview of the thesis is depicted in Fig. 1.1. The four challenges above do not exist in isolation; rather, each corresponds to a key bottleneck in a broader pipeline for converting sparse visual input into an animatable and disentangled 3D human representation. Accordingly, the four subsections below are not independent works, but successive stages that address these bottlenecks from training foundations, to robust body recovery, to disentangled clothed modeling, and finally to controllable avatar generation. We summarise the proposed approaches as follows.

1.3.1 Benchmarking 3D Human Pose and Shape Estimation beyond algorithms (Chapter 3).

To address the lack of systematic benchmarking beyond algorithmic design, we present a comprehensive study of human mesh recovery from three key perspectives:

- **Datasets:** We perform extensive evaluations on 31 datasets, including some that have not been used for mesh recovery before.

- **Backbone:** Beyond the traditional CNN-based feature extractors, our study evaluates 10 backbone architectures, including vision transformers. We further analyze the role of pretraining, showing that initializing weights from a strong pose estimation network provides substantial benefits for mesh recovery.
- **Training strategy:** We investigate different data augmentation and training loss formulations.

1.3.2 Towards Robust and Expressive 3D Human Pose and Shape Estimation (Chapter 4).

To address the robustness limitations in existing whole-body pose and shape estimation (HPSE) methods, we identify three key challenges: (1) inaccurate localization of the subject and its body parts, (2) poor extraction of discriminative and invariant features, and (3) poor image-model alignment, where the projected 3D mesh lacks reprojection consistency with the image.

Motivated by these challenges, we propose a modular framework composed of three novel components, each targeting one of the identified issues:

- **Localization Module**. This module integrates both sparse and dense prediction branches, enabling the model to capture not only the spatial positions but also the semantic context of each body part in the image. By explicitly learning joint locations, it supports the estimation of their relative rotations.
- **Contrastive Feature Extraction Module**. This module adds a pose- and shape-aware contrastive loss, together with positive samples, to enhance feature extraction under challenging augmentations. Positives are formed by image- and location-variant augmentations of the anchor (e.g., color jitter, occlusion, translation, scaling); pose-altering transformations (rotation, flipping) are excluded. Negatives are taken from the rest of the mini-batch, with each anchor–negative pair weighted by its ground-truth pose distance so similar poses are not pushed apart. Minimizing this loss encourages the model to produce stable and consistent representations for the same subject under augmentation, yielding meaningful invariant features.

- **Pixel Alignment Module**. This module leverages differentiable rendering to ensure accurate pixel-level alignment between the projected mesh and the input image, leading to more precise estimation of pose, shape, and camera parameters.

1.3.3 Disentangled Human Generation and Animation from a Single Image (Chapter 5).

To address the problem of entangled body-clothing representations, we propose Disco4D, a method that separates body and clothing modeling using a layered Gaussian Splatting framework where the body is reconstructed with a SMPL-X mesh and embedded Gaussians and the clothing is modeled as layered Gaussians attached to the body surface, with semantic labels for asset-level control.

Our contributions are as follows:

- **Disentangled representation of clothing and human Gaussians**. We design a novel method that separates the reconstruction processes of clothing from human body, enhancing the accuracy and detail of both components.
- **Categorization and Separability of Clothing**. Our method allows for the segmentation of clothing into distinct categories, such as shirts, skirts, shoes and hair. This categorization facilitates the independent manipulation and detailed representation of each clothing item.
- **Animation**. Our methods can enhance the animatability, where the body animations are represented by SMPL-X deformations while clothing deformations are modelled by both body deformation and learned deformations.

1.3.4 Using Video Diffusion Models to Generate Animatable and Customizable Human Avatars with Reusable Assets (Chapter 6).

To mitigate the lack of accessible multi-view canonical supervision and enable animatable avatar generation from a single image, we propose `ReposeHuman`, a video-diffusion-based generation framework:

- We train a **video diffusion model** for dense multi-view image generation in arbitrary poses. The model supports diverse body shapes and sizes, and produces riggable avatars even from partial views. To improve identity consistency and pose control, we adapt two existing conditioning mechanisms to our setting: Image Prompt Adapters (IPA) [35], which we instantiate as separate *Face IPA* and *Clothing IPA* modules for region-specific identity preservation, and Low-Rank Adaptation (LoRA) [36], which we use as a *Pose LoRA* to inject pose-control signals into the diffusion transformer. Our contribution here is the task-specific composition and training of these adapters rather than the adapter mechanisms themselves.
- We curate a **disentangled 3D human dataset** for independent asset control and structured pose conditioning.
- We enhance the **Gaussian Splatting pipeline** for geometrically accurate, high-fidelity asset recovery. We combine 2D Gaussian Splatting, Gaussian Grouping, and normal supervision to reconstruct smooth and clean clothing meshes from disentangled Gaussians.

We demonstrate that `ReposeHuman` generates high-quality, customizable avatars suitable for animation and virtual environments. It provides a scalable multi-view generation solution from a single image, ensuring pose consistency and asset editing. By addressing animation fidelity and disentanglement, `ReposeHuman` improves realism and adaptability for interactive and immersive applications.

1.4 Outline

The thesis is organized into seven chapters, outlined as follows.

Chapter 2 reviews the field of 3D human modelling in two aspects: (1) human pose and shape estimation, and (2) disentangled human generation and animation. These two threads form the conceptual foundation for the technical contributions of this thesis.

Chapter 3 provides a comprehensive benchmarking study for human pose and shape estimation methods beyond algorithmic novelty. In particular, we investigate

the impact of datasets, backbone architectures, and training strategies on model performance. This study establishes strong baselines and offers practical guidance for effective model training and fair evaluation. These findings motivate the architectural developments in Chapter 4.

Chapter 4 introduces `RoboSMPLX`, a robust framework for whole-body 3D human pose and shape estimation. We focus on improving model resilience under real-world conditions by addressing location-variant sensitivity through three key modules: localization awareness, contrastive feature learning, and pixel-level alignment. This chapter strengthens the body-level foundation for the subsequent move to clothed and disentangled avatar generation in Chapter 5.

Chapter 5 presents Disco4D, a novel method for 4D human generation and animation from a single image. This chapter introduces a layered representation that models clothing as a separate Gaussian Splatting-based layer over a canonical SMPL-X body, allowing for detailed, dynamic, and reusable avatar generation. This layered representation motivates the need for canonical multi-view supervision, addressed in Chapter 6.

Chapter 6 proposes ReposeHuman, a video diffusion-based framework for generating animatable 3D human avatars from a single image. The method first synthesizes multi-view canonical poses using 2D video diffusion models, then reconstructs high-quality, disentangled 3D assets. To support this approach, we also introduce a large-scale dataset of animatable 3D meshes.

Chapter 7 concludes the thesis and discusses a few possible future directions.

Chapter 2

Literature Review

This chapter reviews recent research efforts relevant to the four core works in this thesis, covering three primary areas: (1) 3D human pose and shape estimation, and (2) Disentangled 3D human generation and (3) 4D animation. The discussion highlights the limitations of existing methods, which in turn motivate our contributions.

2.1 3D Human Pose and Shape Estimation

2.1.1 Factors Beyond Algorithms

Unlike traditional human pose estimation, which represents the human body using skeletal joint keypoints [37], parametric human models offer a mesh-based representation by encoding human pose and shape through a compact set of parameters. These models typically rely on linear blend skinning and are widely used across various applications due to their efficiency and expressiveness. Notable examples include SMPL [1], SMPL-X [5], STAR [38], and GHUM [39].

Parametric models are capable of producing detailed 3D human meshes with relatively few parameters. For instance, SMPL represents body pose using $\boldsymbol{\theta} \in \mathbb{R}^{72}$ and body shape using $\boldsymbol{\beta} \in \mathbb{R}^{10}$, allowing the reconstruction of a mesh with 6890 vertices. SMPL-X [5] extends SMPL [1] by integrating FLAME [40] for facial

TABLE 2.1: **Summary of the datasets used in various mesh recovery methods and their reported performance (PA-MPJPE in *mm*) on 3DPW and H36M datasets**. Abbreviation for the dataset - Human3.6M [43]: H36M, MPI-INF-3DHP [44]: MI, MuCo-3DHP [50]: MuCo, PoseTrack [51]: PT, OCHuman [52]: OCH. 3DPW *Protocol 2* (P2) refers to the evaluation (PA-MPJPE) on 3DPW test set without training on 3DPW train set while *Protocol 1* (P1) includes fine-tuning on 3DPW train set. We use the notation $[*]_{\text{EFT/ SPIN/ DP/ SMPLify-X}}$ to denote datasets with EFT, SPIN, DensePose or SMPLify-X fittings.

Method	Datasets used	Backbones	Losses	3DPW (P2)↓	3DPW (P1)↓	H36M↓
HMR [2]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	Mixed	76.7	-	56.8
NBF [53]	H36M, UP-3D, HumanEva-I	ResNet-50	Mixed	-	-	59.9
GraphCMR [10]	H36M, UP-3D, COCO, LSP, MPII	ResNet-50	Mixed	70.2	-	-
HoloPose [9]	H36M, MPII, $[\text{COCO}]_{\text{DP}}$	ResNet-50	-	-	-	46.5
SPIN [3]	H36M, $[\text{MI}]_{\text{SPIN}}$, COCO, LSP, LSPET, MPII	ResNet-50	Mixed	59.2	-	41.1
[11]	H36M, MI, PT, LSP, LSPET, MPII, COCO	ResNet-50	-			52.7
[54]	H36M, $[\text{COCO}]_{\text{DP}}$, UP3D, $[\text{LSP, LSPET, MPII, COCO}]_{\text{SPIN}}$	-	-			41.7
Pose2Mesh [12]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO, Freihand	PoseNet	-	58.9		47
HKMR [55]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	L1			43.2
I2L-MeshNet [56]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO, Freihand	ResNet-50	-	57.7		41.1
DaNet [57]	H36M, $[\text{COCO}]_{\text{DP}}$, UP3D, $[\text{LSP, LSPET, MPII, COCO}]_{\text{SPIN}}$	-	-	54.8		40.5
Pose2Pose [13]	MuCo, $[\text{H36M}]_{\text{SMPLify-X}}$, COCO-Wholebody, Freihand	-	-	55.3		47.4
HybrIK [58]	H36M, MI, COCO	ResNet-34	-	48.8		
METRO [59]	H36M, UP-3D, MuCo, COCO, MPII, Freihand	HRNet-W64	-		47.9	36.7
BMP [60]	H36M, MI, MuCo, COCO, LSP, LSPET, PT, MPII	ResNet-50	MSE	63.8		51.3
HUND [39]	H36M, 3DPW, COCO-2017, OpenImages	-	-	57.5		53
EFT [14]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$	ResNet-50	Mixed	54.2	52.2	
ProHMR[15]	H36M, $[\text{MI, COCO, MPII}]_{\text{SPIN}}$	ResNet-50	Mixed		59.8	41.2
DSR [16]	H36M, MI, $[\text{COCO}]_{\text{EFT}}$	ResNet-50	Mixed	54.1	51.7	
ROMP [61]	H36M, UP-3D, $[\text{MI, COCO, MPII, LSP}]_{\text{SPIN}}$, AICH	ResNet-50	-	54.9	62	
ROMP[61]	H36M, UP-3D, $[\text{MI, COCO, MPII, LSP}]_{\text{SPIN}}$, AICH, PT, CrowdPose, MuCo, OH	ResNet-50	-	53.3	56.8	
Graphormer [62]	H36M, MuCo, UP-3D, COCO, MPII	HRNet-W64	L1		45.6	34.5
THUNDR [63]	H36M, 3DPW, COCO-2017, OpenImages	ResNet-50	-	51.5		39.8
PyMAF [64]	H36M, $[\text{MI}]_{\text{SPIN}}$, COCO, LSP, LSPET, MPII	ResNet-50	-	58.9	51.2	40.5
SPEC [17]	Pano360, SPEC-SYN, SPEC-MTP, 3DPW, MI, H36M, $[\text{COCO, MPII, LSPET}]_{\text{EFT}}$	ResNet-50	-	53.2		
PARE [4]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$, MI, H36M	ResNet-50	Mixed	52.3		
PARE [4]	$[\text{COCO, MPII, LSPET}]_{\text{EFT}}$, MI, H36M	HRNet-W32	Mixed	50.9	46.5	

modeling and MANO [41] for hands, and is trained on large-scale 3D scan data to capture diverse human shapes.

For a broader overview of monocular 3D human mesh reconstruction, Tian et al. [42] surveyed various model architectures and benchmark results. They highlighted factors such as output type, pseudo supervision, dataset selection, and evaluation protocols as sources of performance variance, though no empirical validation was provided. In contrast, our work conducts systematic experiments to analyze how dataset choice, network architecture, and training strategy influence model performance, providing practical insights for robust HPSE.

Datasets. Kanazawa et al. [2] combined Human3.6M (H36M) [43], MPI-INF-3DHP [44], COCO [45], LSPET [46], LSP [47] and MPII [48]. To utilize multiple datasets,they concatenated the datasets using a manually defined sampling ratio, ensuring that datasets with a substantially larger number of samples (i.e. H36M) do not dominate the training process [2, 3]. Recently, more competitive datasets are introduced [14, 49] for training high-performing models.

Table 2.1 summarises the datasets used in various human mesh recovery algorithms.

Many existing algorithms are trained on distinct combinations of datasets, yet their performance on the 3DPW [65] test set is often compared directly, even when the training data differ.

Further complicating comparisons, [39] highlighted that multiple evaluation protocols have been used. Following SPIN [3], most studies adopt protocol 2, in which the 3DPW test set is evaluated without fine-tuning on the 3DPW training split. However, some works instead apply protocol 1, incorporating the 3DPW training set during model training [8, 17, 59, 66–68]. For H36M [43], at least four protocols exist: the original evaluation protocol defined by the dataset creators, evaluation on the withheld test set, and protocols 1 and 2 introduced by [3], which repartition the original training and validation sets with available ground truth. More recently, [39] added evaluation on Panoptic-test [69] and MuPoTs-3D-test [50], while [70] recommended AGORA-test and [14] proposed EFT-OCHuman-test and EFT-LSPET-test for more challenging benchmarks.

Architectures.

Since HMR proposed by Kanazawa et al. [2], ResNet-50 [71] has been the default backbone in many mesh recovery pipelines [3, 4, 7, 14]. More recently, PARE [4] replaced ResNet-50 with HRNet-W32 [72], attributing the observed performance improvements to HRNet-W32's ability to generate high-resolution, robust feature representations. Cai et al. [73] further explored alternative backbones, including deeper CNNs such as ResNet-101 and ResNet-152 [71], as well as DeiT [74], a vision transformer. As expected, larger-capacity models generally perform better [73], though vision transformers did not show consistent advantages over CNN-based architectures.

Training strategy. Following HMR [2], most mesh recovery frameworks use a combination of losses: Mean Squared Error (MSE) for keypoint supervision and L1 loss for supervising SMPL parameters.

Augmentation techniques widely used in pose estimation [3–5, 14, 50, 55, 60, 75–78] have also been adopted for mesh recovery, though with varying effectiveness. Occlusion augmentation has been shown to provide substantial benefits [55, 60], while other works report only minor gains [7]. Joo et al. [14] report that extreme cropping offers only marginal improvement, and PARE [4] found that it can even reduce performance on the 3DPW benchmark. Since these findings are based on

different dataset compositions and benchmarks, a more controlled and systematic investigation is needed to assess the impact of augmentation on individual datasets.

2.1.2 Algorithmic Robustness

Whole-body Mesh Recovery. Despite significant progress in 3D body-specific [3, 4, 10, 12, 14–17], hand-specific [59, 79], and face-specific [80] mesh recovery methods, there have been limited attempts to simultaneously recover all those parts. Early studies on whole-body pose and shape estimation primarily fit a 3D human model to 2D or 3D evidence [5, 81–83], which can be slow and susceptible to noise. Recent studies utilized neural networks to regress the SMPL-X parameters for a whole-body 3D human mesh. The model is composed of separate sub-networks to process body, hand and face, respectively. *One-stage* methods, e.g., OS-X [84], have the benefit of reduced computational costs and improved communication within part modules for more natural mesh articulation. However, the omission of hand and face experts makes it difficult for the model to leverage the widely available part-specific datasets, thus decreasing the hand and face performance. *Multi-stage* methods, e.g., ExPose [85], FrankMocap [86], PIXIE [87] and Hand4Whole [13], use different techniques to localize part crops.

In response, recent studies utilized neural networks to regress the SMPL-X parameters for a whole-body 3D human mesh. The model is composed of separate sub-networks to process body, hand and face, respectively. These methods can be classified into two categories. (1) *One-stage* methods, e.g., OS-X [84], have the benefit of reduced computational costs and improved communication within part modules for more natural mesh articulation. However, the omission of hand and face experts makes it difficult for the model to leverage the widely available part-specific datasets, thus decreasing the hand and face performance. (2) *Multi-stage* methods, e.g., ExPose [85], FrankMocap [86], PIXIE [87] and Hand4Whole [13], use different techniques to localize part crops. Expose [85] and PIXIE [87] localize hand and part crops from the body mesh, making them dependent on the accuracy of body poses. Minor rotation errors accumulated along the kinematic chain may result in deviations in joint locations and thus inaccurate part crops. In contrast, Hand4Whole [13] predicts hand and face bounding boxes using a network leveraging image features and 3D joint heatmaps, but the resulting crops have low

resolution. PyMAF-X [57] relies on an off-the-shelf whole-body pose estimation model to obtain crops, which, while more accurate, is not an end-to-end solution.

Contrastive Learning. Contrastive learning has recently emerged as a leading paradigm in self-supervised learning, achieving strong performance in various domains. This approach has been applied to 3D hand pose and shape estimation [88, 89].

In the context of 3D human pose and shape estimation, [90] was the first to investigate its utility, reporting that self-supervised contrastive learning offered limited benefits because the resulting embeddings were difficult to associate with high-level human-specific features. Supervised contrastive learning, introduced in [91] for image classification, integrates label information to bring embeddings of the same class closer and push apart those from different classes. However, no prior work has adapted this strategy for human pose and shape estimation, where defining positive samples is challenging and the data lies in a continuous space. Our work is the first to address these challenges, applying supervised contrastive learning to whole-body pose and shape estimation.

Pixel Alignment in Pose and Shape Estimation. Numerous methods aim to improve the localization of subjects in images. Some approaches rely on implicit supervision, projecting the predicted mesh and supervising the corresponding 2D joints [2–4, 10, 12, 14–17]. Others enhance alignment with additional dense cues, such as body landmarks, silhouettes, or part segmentation [4, 16, 53, 54, 66, 78, 92]. Alternatively, explicit localization methods directly predict the positions of body parts in the image. For instance, [13] predicts keypoint coordinates, while semantic body part segmentation has been used as an intermediate representation [4, 53]. PARE [4] employs differentiable rendering to project the ground-truth mesh into image space and supervise part silhouettes. However, dense prediction and differentiable rendering have yet to be fully integrated into whole-body pose and shape estimation. A complementary direction is taken by Neural Localizer Fields [93], which reformulate pose and shape estimation as a continuous neural field over the human body volume, enabling localization of arbitrary queried 3D points in the image. This unifies heterogeneous supervision (e.g., mesh, 2D/3D skeleton, dense pose) without format conversion and decouples representation from specific parametric outputs. In contrast, our framework remains within the parametric regression paradigm and targets a different bottleneck—jointly

improving localization, feature robustness under input perturbations, and image-space consistency.

2.2 3D Disentangled Human Generation

3D human generation aims to generate 3D human representations from image or text conditions. This section first reviews representations for 3D human generation.

Table 2.2 summarizes the relevant 3D/4D generation methods. We describe their details below.

2.2.1 3D Human Generation

Single-image 3D Generation. Single-image reconstruction leverages advanced methods [94, 95] to generate 3D assets in the form of 3D point clouds or NeRF [96] from one image. While earlier efforts using auto-encoders focused on synthetic objects [78, 97–101], newer approaches treat the task as conditional generation, employing diffusion models [102] for 3D generation from both image and text [102–109]. One-2-3-45 [110] uses 2D diffusion models [105, 111] to generate multi-view images for reconstruction, while LRM [112] adopts transformer-based architecture to scale up the task on large datasets [106, 113]. Gaussian-based methods [114], particularly DreamGaussian [115] and LGM [116], offer efficient, high-resolution 3D model generation from text or images. Recently, video diffusion models have attracted significant attention due to their remarkable ability to generate intricate scenes and complex dynamics with great spatio-temporal consistency [117–123]. They are employed to generate consistent multi-view images, and then reconstruct underlying 3D assets with high quality [124].

Single-image human-centric 3D Generation. Significant research efforts have been made for 3D human reconstruction, which can be classified into the following categories. (1) *Explicit-shape-based methods* rely on Human Mesh Recovery (HMR) using parametric models like SMPL [1] and SMPL-X [5] to generate 3D body meshes [2–4, 12–14, 16, 17, 57, 85–87]. To account for 3D garments, several approaches incorporate offsets [125, 126] or templates, utilize deformable garment templates [127, 128], or employ non-parametric forms for clothed figures [24, 129, 130]. Despite

TABLE 2.2: **3D/4D generation methods from a single image.**

Method	Type	Layered	Animatable
LGM [116]	General	✗	✗
PiFU [19]	Human-centric	✗	✗
DreamFusion [108]	General	✗	✗
DreamGaussian [115]	General	✗	✗
PiFU [19]	Human-centric	✗	✗
D-IF [132]	Human-centric	✗	✗
HiLo [133]	Human-centric	✗	✗
ECON [24]	Human-centric	✗	✗
SHERF [26]	Human-centric	✗	✓
Disco4D	Human-centric	✓	✓

their advancements, they face limitations in handling complex outfit variations and loose clothing due to inherent topological constraints. (2) *Implicit-function-based methods* utilize implicit representations like occupancy or distance fields for modeling clothed humans with complex geometries, such as loose garments. Techniques range from end-to-end regression of free-form implicit surfaces [18–20] to use of geometric priors [21–25] and implicit shape completion [24]. Notable works such as PIFu [19], ARCH(++) [21, 22], and PaMIR [23] can extract textured models from images, but struggle with depth ambiguities and texture inconsistencies. (3) *NeRF-based methods* incorporate model-based priors (i.e., SMPL-X) for accurate human reconstruction. Efforts like SHERF [26] and ELICIT [27] improve the reconstruction coherence by addressing 2D observation incompleteness leveraging appearance priors. (4) *Diffusion-plus-Gaussian methods* combine multi-view diffusion priors with 3D Gaussian Splatting, using the explicit 3D representation to enforce cross-view consistency. Human3Diffusion [131], for example, couples a multi-view diffusion model with a Gaussian reconstruction branch, enabling the two components to co-adapt during training and produce high-fidelity geometry and appearance from a single image. However, these methods typically yield entangled human representations, limiting fine-grained control over individual components such as clothing, hair, and accessories, motivating the disentangled representations explored in this thesis.

3D Clothing Modeling. Reconstructing clothing from images and videos as a separate layer over the human body poses significant challenges due to the diversity of clothing topologies. Previous efforts relied on either template meshes or implicit surface models, and required extensive, high-quality 3D data from simulations [134–137] or tailored template meshes [138–141]. New methods were developed [127, 142]

for multi-clothing models and versatile template meshes, respectively, facilitating diverse clothing topology encoding. However, these techniques typically fall short in capturing the clothing texture and appearance. The reliance on predefined clothing style templates further constrains their ability to handle real-world clothing variations. Corona et al. [143] addressed these shortcomings by representing clothing layers with deep unsigned distance functions and an auto-decoder for style and cut differentiation, though this often produces overly-smooth reconstructions [143]. On the other hand, SCARF [144] and DELTA [145] significantly enhance the visual fidelity by applying NeRF to clothing layers, but require self-rotating video inputs and considerable processing times.

Multi-view Canonical Image Synthesis. Inspired by these successes [105, 111], newer approaches apply diffusion to multi-view human image generation for 3D human reconstruction [146, 147]. However, these models remain static and non-animatable, limiting their applicability for dynamic avatar creation and real-time animation.

Diffusion models achieve strong quality and control for image-based human animation [28–30, 32, 148, 149]. Many use a UNet-based ReferenceNet and pose guider for pose-aware synthesis. CHAMP [32] adds denser controls by combining SMPL-rendered normals, semantic maps, depth, and skeleton guidance in its latent model. However, these models rely on 2D video data and lack 3D supervision. As a result, they suffer from dynamic clothing artifacts, poor temporal consistency, texture incoherence, inconsistent facial features, and body shape changes across views. Their inability to generate consistent multi-view images limits their effectiveness for 3D avatar generation.

For animatable avatars, CharacterGen [33] and EN3D [34] generate riggable models from four A-pose images. However, their reposing often fails to preserve fine body shape, facial details, and clothing accuracy. Modeling in A-pose also limits articulation and causes artifacts from self-occlusion during reposing or novel view synthesis. Using T-pose, X-pose, or DA-pose offers more varied limb orientations, reduces self-occlusion, and improves realism, generalization, and adaptability.

3D Clothed Human Datasets. High-precision 3D humans are usually created by scans or multi-view capture. Scan datasets like THuman2.0 [150], Twindom, and 2K2K [151] capture detailed geometry but cover few subjects and simple poses. Large

datasets like Objaverse [106] and MVImgNet [113] focus on general objects and lack human-specific details. Diffusion-based 2D datasets such as HuGe100K [152] improve pose, body shape, and clothing diversity but may introduce body inconsistencies, artifacts, and facial identity variations, lowering their reliability for training.

2.3 4D Human Animation

4D Animation. This task aims at capturing dynamic 3D scenes over time. Two primary approaches have emerged: modeling 4D scenes by adding time dimension t or latent codes to spatial coordinates [153–155]; combining deformation fields with static 3D scenes [156–162]. Recent efforts in explicit or hybrid representations, like planar decomposition [163–165], hash representations [166], and other innovative methods [167–169], have improved reconstruction speed and quality. Gaussian Splatting, especially, stands out for balancing efficiency with quality, with dynamic 3D Gaussians [170] and 4D Gaussian Splatting [171, 172] techniques introducing time-dependent deformations to enhance reconstructions. Notably, DreamGaussian4D [173] stands out by minimizing the optimization time while achieving high-quality 4D reconstructions.

Human-centric 4D Animation. Recent works leverage Gaussian-based methods [174–180] for 4D human reconstruction, requiring extensive frame sequences (50-100 frames) and/or multiple viewpoints. Currently there has not been any work on 4D layered human generation and animation from a single image or a video with few images, which will be achieved in this paper.

2.4 Conclusion

The literature reviewed in this chapter reveals four key gaps, each motivating one of the technical contributions of this thesis.

Gap 1 (Benchmarking rigor). Progress in human mesh recovery has been reported under widely differing dataset compositions, backbone architectures, and training protocols, making cross-method comparisons difficult to interpret. There has been limited systematic study that holds the algorithm constant while varying

these factors, which is necessary to establish fair baselines and to quantify the contribution of each component.

Gap 2 (Robustness of whole-body estimation). Existing whole-body methods typically rely on separate, independently cropped sub-networks for body, hand, and face, and therefore depend heavily on bounding-box quality. Empirical evidence suggests that even small perturbations in crop scale or alignment can lead to significant error increases. This highlights three underexplored challenges: accurate subject localization, learning feature representations that are invariant to input perturbations, and achieving consistent image–model alignment.

Gap 3 (Disentangled single-image 3D human generation). Beyond the commonly cited scarcity of paired clothed-human data, disentangled single-image generation is constrained by several structural limitations. First, many methods represent body and clothing as a single fused mesh, limiting asset extraction and fine-grained editing. Second, representations that tightly couple geometry and appearance make it difficult to attach semantic meaning to individual components such as clothing. Third, occluded regions are often inferred without strong geometric constraints, leading to inconsistencies in reconstruction. These limitations suggest that data scarcity alone does not fully explain the challenges in this setting.

Gap 4 (Animatable, customizable avatars from a single image). Current approaches either require impractical capture conditions, such as self-rotating videos, or rely on single-image diffusion models trained with weak 2D supervision, which often fail to preserve body shape and struggle with occlusions. Methods that synthesize canonical multi-view images provide only sparse coverage, limiting geometric accuracy, and generally do not support disentangled, reusable representations of clothing and other assets.

These gaps directly motivate the contributions of this thesis. Specifically, the subsequent chapters address benchmarking rigor (Chapter 3), robustness in whole-body estimation (Chapter 4), disentangled 3D human generation (Chapter 5), and the creation of customizable, animatable avatars from minimal input (Chapter 6).

Chapter 3

Benchmarking 3D Pose and Shape Estimation Beyond Algorithms

3.1 Introduction

This chapter addresses the first stage of the broader pipeline shown in Fig. 1.1. Rather than proposing a new architecture, we ask what actually drives human mesh recovery performance in practice. The findings of this chapter establish the foundation for the subsequent work on robust whole-body pose and shape estimation in Chapter 4.

Despite significant progress in mesh recovery algorithms, previous studies have rarely conducted systematic analyses of other fundamental factors that critically influence model performance. (1) Different choices of datasets and their respective contributions can lead to varying results. This is particularly evident in human mesh recovery, where datasets containing different label modalities (e.g., 2D keypoints, 3D keypoints, segmentation masks, SMPL parameters) are often combined for training. (2) Mesh recovery models are commonly trained using a pretrained backbone, and the quality of this backbone—such as its network architecture and weight initialization—is a key determinant of downstream performance. (3) Model performance is also highly sensitive to training strategies, including data augmentation techniques and loss function design. *It is still unclear how these factors*

The work in this chapter has been published in [181].

can affect the model performance and what are the optimal training configurations to obtain good mesh recovery models.

This lack of understanding can significantly hinder the advancement of mesh recovery research. In particular, new algorithms may be developed and evaluated under suboptimal training settings, preventing their full potential from being accurately reflected. For instance, the state-of-the-art algorithms SPIN [3] and PARE [4] can achieve the PA-MPJPE (*i.e.*, recovery error) of 59.2 ***mm*** and 50.9 ***mm***, respectively, while we can obtain the PA-MPJPE of 47.3 ***mm*** by selecting a better configuration with a simple base method (Table 3.1). Second, several previous studies have compared algorithms or methods that were trained under different configurations, which can result in biased or inconsistent evaluations. For instance, HMR [2] and SPIN [3] are commonly adopted as baselines in comparisons with a range of other approaches [4, 14, 16, 17, 58], even though they rely on markedly different combinations of training datasets. In contrast, only a limited number of works [17, 64] have ensured fairness by using the identical dataset composition as HMR or SPIN, or by replicating HMR's dataset configuration for ablation experiments.

To tackle the issues outlined above, we conduct an extensive benchmarking study on human mesh recovery from three key perspectives. **(1) Datasets.** We carry out thorough evaluations across 31 datasets, including several that have not previously been explored for mesh recovery. Our findings show that careful dataset selection can yield substantial performance improvements. We highlight the characteristics that make a dataset effective and offer recommendations for improving existing datasets or designing new ones. **(2) Backbone.** While most existing methods still employ traditional CNN-based feature extractors [4, 16], our study expands the comparison to 10 different backbone architectures, incorporating vision transformers. We further analyze the role of pretraining and find that initializing weights from a high-performing pose estimation model significantly benefits mesh recovery performance. **(3) Training strategy.** We investigate a range of data augmentation techniques and loss function designs, revealing that L1 loss provides better supervision and noise suppression than the commonly used mixed-loss formulations. We also explain the varying effectiveness of different augmentation strategies in terms of the feature distribution discrepancies between training and testing datasets.

Bringing together our findings, we establish strong baselines for various dataset combinations and backbones using the HMR framework [2] and evaluating on the

TABLE 3.1: **Our identified optimal baseline models with the performance on the 3DPW test set.** Abbreviations for the datasets - Human3.6M [43]: H36M, MPI-INF-3DHP [44]: MI, MuCo-3DHP [50]: MuCo, PoseTrack [51]: PT, OCHuman [52]: OCH

Algorithm	Dataset	Backbone	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
PARE [4]	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	50.90	82.0	-	97.9
Ours	EFT-[COCO, LSPET, MPII], H36M-Aug, SPIN-MI	HrNet-W32	47.68	81.16	64.70	98.23
SPIN [3]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	59.2	96.9	-	135.1
HMR [2]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	76.7	130.0	-	-
Ours	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	51.66	82.80	70.53	100.59
Ours	H36M, MI, COCO, LSP, LSPET, MPII	Twin-SVT-B	48.77	82.91	66.91	96.33
Ours	H36M, MI, COCO, LSP, LSPET, MPII	HrNet-W32	49.18	79.76	68.58	96.07
Ours	H36M-Aug, MI, COCO, LSP, LSPET, MPII	Twin-SVT-B	47.70	**79.16**	66.53	**95.03**
Ours	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twin-SVT-B	**47.31**	81.90	**64.19**	96.56
Ours	H36M, MI, EFT-COCO	HrNet-W32	48.08	83.16	66.01	100.59
Ours	H36M, MI, EFT-COCO	Twin-SVT-B	48.27	84.39	64.72	99.61
Ours	H36M, MuCo, EFT-COCO	Twin-SVT-B	47.76	80.03	64.43	98.07
Ours	EFT-[COCO, LSPET, PT, OCH] H36M, MI	Twin-SVT-B	49.33	83.13	66.29	99.73

3DPW test set [65], as summarized in Table 3.1. According to [70], performance on the 3DPW-test benchmark appears to be reaching saturation in the PA-MPJPE range of 50+*mm*, making it harder to gauge progress toward fully robust and generalizable solutions. In this work, we achieve a PA-MPJPE of 47.68 *mm* using the same backbone and dataset selection as PARE [4], which reports 50.9 *mm* despite employing a more complex design. When matching the model capacity and dataset selection of HMR (76.7 *mm*) [2] and SPIN (59.2 *mm*) [3], our approach attains 51.66 *mm*. Using HMR's original dataset and split—without any EFT or SPIN fittings—we further achieve 48.77 *mm*.

With improved dataset selection inspired by [4], our best-performing configuration attains 47.31 *mm* without fine-tuning on the 3DPW training set. We hope these competitive results encourage the community to prioritize advancements in algorithmic design over variations in training setups in future research.

3.2 Preliminaries

Base model. The origin of many mesh recovery works [3–5, 7, 8, 15, 16, 182] can be traced back to HMR [2]. HMR employs a neural network to directly regress the parameters of the SMPL body model [2], a differentiable function mapping pose parameters θ and shape parameters β to a triangulated mesh consisting of 6,980 vertices. Building upon this foundation, later studies have introduced modifications to improve accuracy and robustness. Examples include incorporating an optimization loop [3], estimating camera parameters [17], or adopting probabilistic pose estimation [15]. Other extensions expand HMR to predict additional dimensions, such as appearance (e.g., HMAR [183]) or temporal information (e.g., HMMR [184],

VIBE [7], MEVA [8]). We select HMR as a benchmark model given its extensive use as a baseline in prior work [14, 16, 70, 73]. Section 3.6 further reports benchmarking results for additional algorithms.

Evaluation. We adopt the standard evaluation protocol described in [2, 3], using PA-MPJPE (in mm) as the primary metric, where lower values indicate more accurate recovery. The aim is to estimate pose parameters θ and shape parameters β, which are then passed to the parametric human model to produce joint coordinates. This metric inherently captures both shape and mesh accuracy [3, 42, 59, 62]. Since PA-MPJPE alone may not fully reflect performance [64, 185], we also report additional measures including PVE, PA-PVE, and MPJPE.

Our main evaluation is conducted on the 3DPW [65] test set without fine-tuning on its training data (*Protocol 2*)[0]. Section 3.6 also provides results on additional benchmarks, reinforcing that 3DPW is representative for in-the-wild evaluation.

Because 3DPW contains a diverse range of outdoor scenes, it is frequently used as the primary or sole benchmark for assessing real-world performance [2–4, 7, 14, 16, 61]. For completeness, we also evaluate on the indoor H36M [43] test set. The results, shown in Table 3.2, are consistent with those obtained on 3DPW.

All models are trained for 100 epochs[1] and evaluated at each epoch, with the best PA-MPJPE reported. Our benchmarking is carried out from three perspectives – datasets (Section 3.3), backbones (Section 3.4), and training strategies (Section 3.5).

3.3 Benchmarking Training Datasets

Training datasets are a critical factor in determining mesh recovery performance. Table 2.1 in Section 2 summarises the datasets employed across different algorithms. Many studies rely on their own heuristically chosen dataset combinations [4, 14, 16, 17, 58], making it difficult to disentangle performance improvements due to the proposed method from those arising from curated dataset selection. This highlights the need for controlled benchmarks using varied dataset configurations. In this

[0]Some studies [8, 17, 59, 66–68] use *Protocol 1*, which includes the 3DPW training set. While this generally leads to improved results, it is not consistently adopted across works.

[1]All experiments are conducted using 8 Tesla V100 GPUs.

TABLE 3.2: **HMR model performance (PA-MPJPE in *mm*) on the 3DPW [65] and H36M[43] test sets when trained on individual datasets. For PROX and MuPoTs-3D, only 2D keypoints are used for training. P: person-person occlusion O: person-object occlusion.**

. Extra evaluation on other datasets (AGORA, MPI-INF-3DHP, EFT-COCO, MuPots-3D, EFT-OCH, EFT-LSPET) and other metrics for (MPJPE, PA-PVE, PVE) for 3DPW can be found in the main paper.

Training dataset	Annotation type	Env.	# Samples	# Subjects	# Scenes	# Cam	Occ.	3DPW↓	H36M↓
PROX [186] *	2DKP	Indoor	88484	11	12	-	O	84.69	112.31
COCO-Wholebody [191]	2DKP	Outdoor	40055	40055	-	-	-	85.27	95.51
Instavariety [6]	2DKP	Outdoor	2187158	>28272	-	-	-	88.93	98.74
COCO [45]	2DKP	Outdoor	28344	28344	-	-	-	93.18	97.72
MuPoTs-3D [50] *	2DKP	Outdoor	20760	8	-	12		95.83	137.60
LIP [187]	2DKP	Outdoor	25553	25553	-	-	-	96.47	113.09
MPII [48]	2DKP	Outdoor	14810	14810	3913	-	-	98.18	121.46
Crowdpose [188]	2DKP	Outdoor	13927	-	-	-	P	99.97	123.47
Vlog People [6]	2DKP	Outdoor	353306	798	798	-	-	100.38	121.42
PoseTrack (PT) [51]	2DKP	Outdoor	5084	550	550	-	-	105.30	135.05
LSP [47]	2DKP	Outdoor	999	999	-	-	-	111.45	153.36
AI Challenger [189]	2DKP	Outdoor	378374	-	-	-	-	111.66	115.07
LSPET [46]	2DKP	Outdoor	9427	9427	-	-	-	112.26	125.44
Penn-Action [190]	2DKP	Outdoor	17443	2326	2326	-	-	114.53	130.17
OCHuman (OCH) [52]	2DKP	Outdoor	10375	8110	-	-	P,O	130.55	131.77
MuCo-3DHP (MuCo) [50]	2DKP/ 3DKP	Indoor	482725	8	-	14	P	78.05	106.08
MPI-INF-3DHP (MI) [44]	2DKP/ 3DKP	Indoor	105274	8	1	14	-	107.15	132.49
3DOH50K (OH) [54]	2DKP/ 3DKP	Indoor	50310	-	1	6	O	114.48	132.38
3D People [192]	2DKP/ 3DKP	Indoor	1984640	80	-	4	-	108.27	117.19
AGORA [70]	2DKP/ 3DKP/ SMPL	Indoor	100015	>350	-	-	P,O	77.94	105.22
SURRREAL [193]	2DKP/ 3DKP/ SMPL	Indoor	1605030	145	2607	-	-	110.00	149.99
Human3.6M (H36M) [43]	2DKP/ 3DKP/ SMPL	Indoor	312188	9	1	4	-	124.55	52.68
EFT-COCO [14]	2DKP/ SMPL	Outdoor	74834	74834	-	-	-	60.82	72.87
EFT-COCO-part [14]	2DKP/ SMPL	Outdoor	28062	28062	-	-	-	67.81	82.36
EFT-PoseTrack [14]	2DKP/ SMPL	Outdoor	28457	550	-	-	-	75.17	94.74
EFT-MPII [14]	2DKP/ SMPL	Outdoor	14667	3913	-	-	-	77.67	93.77
UP-3D [194]	2DKP/ SMPL	Outdoor	7126	7126	-	-	-	86.92	181.7
MTP [195]	2DKP/ SMPL	Outdoor	3187	3187	-	-	-	87.03	93.69
EFT-OCHUMAN [14]	2DKP/ SMPL	Outdoor	2495	2495	-	-	P,O	94.01	109.85
EFT-LSPET [14]	2DKP/ SMPL	Outdoor	2946	2946	-	-	-	100.53	112.03
3DPW [65]	SMPL	Outdoor	22735	7	-	-	-	89.36	130.63

work, we present a systematic and extensive evaluation of how training datasets influence HMR accuracy. Our benchmarks include not only datasets commonly used in earlier mesh recovery research, but also recently introduced ones (e.g., PROX [186], AGORA [70]) and datasets widely adopted in 2D/3D pose estimation (e.g., LIP [187], CrowdPose [188], AI Challenger [189], Penn-Action [190], MuCo-3DHP [50], etc.). We further examine specific dataset characteristics that can affect model performance—factors that have received little attention in prior work.

3.3.1 Dataset Attributes

Different datasets possess distinct characteristics that can significantly influence model performance. To facilitate a clear analysis of their effects, we individually train the HMR model on each dataset in our collection and then evaluate its performance. The attributes of these datasets, along with the corresponding results, are presented in Table 3.2.

Non-critical attributes. [14] reported an indoor–outdoor domain gap, where models trained on outdoor datasets tend to underperform on indoor benchmarks, and vice versa. However, our large-scale benchmarks show that this is an oversimplification, and performance cannot always be explained solely by the indoor–outdoor

gap. For example, models trained on some indoor datasets (*e.g.*, PROX, MuCo-3DHP) achieve strong results on the outdoor 3DPW benchmark, surpassing many models trained on outdoor datasets. Conversely, models trained on certain indoor datasets (*e.g.*, MPI-INF-3DHP, 3DOH50K) perform poorly on the indoor H36M test set (see Table 3.2). We also observe only weak correlations between dataset size and performance. For instance, COCO, with roughly one-tenth the number of samples in H36M [43], still achieves higher accuracy on the 3DPW test set.

Critical attributes. Several dataset characteristics strongly influence model accuracy, including human pose variety, body shape (height, limb proportions), scene type, lighting conditions, occlusion (self, inter-person, and environmental), annotation type (2D/3D keypoints, SMPL), and camera properties (e.g., viewing angles) [14, 42, 70, 73, 196, 197]. Greater similarity between these attributes in the training and test sets generally leads to better performance.

To investigate this, we use a well-trained HMR model to extract distributions for four attributes: 1) pose $\theta \in R^{69}$, 2) shape $\beta \in R^{10}$ and 3) camera translation $t^c \in R^3$ obtained from the head, and 4) features $f \in R^{2048}$ obtained from the ResNet-50 backbone. Fig. 3.1 presents UMAP visualizations [198] for four representative datasets: COCO, 3DPW, H36M, and MPI-INF-3DHP. We have the following observations. First, COCO exhibits high diversity in all attributes, with significant overlap with 3DPW, explaining its strong performance on that benchmark. Second, H36M shows limited pose diversity (Fig. 3.1a) and has distinctly different distributions of features (Fig. 3.1d) and shape (Fig. 3.1b) from 3DPW. This is likely due to the small number of subjects (9) and a single capture environment (Table 3.2). In addition, H36M's shape and camera distributions also differ from those of MPI-INF-3DHP, which contributes to poor cross-dataset performance between them. Full H36M benchmark results and additional visualizations of attribute distributions for other datasets are provided in [181].

Notably, indoor datasets that achieve strong results on the outdoor 3DPW benchmark often contain a high degree of person–person occlusion (MuCo-3DHP) or person–object occlusion (PROX) (Fig. 3.2). In real-world settings, occlusion frequently occurs due to self-overlapping body parts, close proximity between individuals, or contact with surrounding objects [70]. Our findings suggest that such occlusion plays a more influential role in cross-domain generalization than background

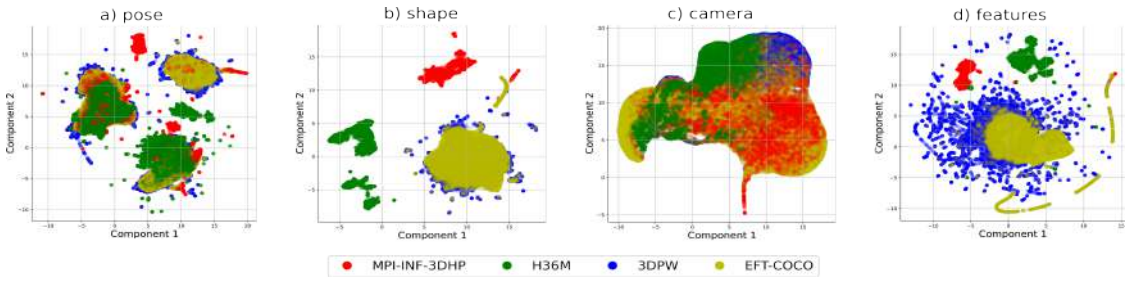


FIGURE 3.1: **Distributions of pose, shape, camera, backbone feature distributions in four datasets (better viewed in color)**.

(a) MPI-INF-3DHP [44] (b) MuCo-3DHP [50] (c) PROX [186]

FIGURE 3.2: [
Example images from MPI-INF-3DHP, MuCo-3DHP, and PROX]**Example images sourced from (a) MPI-INF-3DHP [44] (b) MuCo-3DHP [50] and (c) PROX [186].**

differences. This is evident in two cases: (1) MuCo-3DHP [50], created by compositing MPI-INF-3DHP [44] with inter-person occlusions, leads to substantially better HMR performance than MPI-INF-3DHP alone (78.05 vs. 107.15 PA-MPJPE (*mm*)). (2) PROX [186], the only indoor 2D keypoint dataset containing numerous person-object interactions with furniture (Fig.3.2), achieves the best 3DPW performance (PA-MPJPE of 84.69 *mm*) among all 2D keypoint datasets, including outdoor ones. Importantly, when background, lighting, and actors are held constant, adding inter-person occlusion in MuCo-3DHP leads to significant performance gains. This improvement is further supported by the observation that the pose, shape, camera, and feature distributions of MuCo-3DHP and PROX (see Figure 3.3(a) - 3.3(d)) are closer to those of 3DPW-test [65] and other in-the-wild datasets than MPI-INF-3DHP, highlighting the central role of occlusion in bridging the domain gap.

To highlight the significance of the SMPL fitting process, we compare EFT datasets

TABLE 3.3: **HMR model performance with EFT datasets**.

Dataset	w/ SMPL	w/o SMPL
EFT-COCO	60.82	94.42
EFT-COCO-Part	67.81	101.65
EFT-PoseTrack	75.17	103.10
EFT-MPII	77.66	99.87
EFT-OCHuman	94.01	121.68
EFT-LSPET	100.53	134.62

with and without SMPL annotations, as reported in Table 3.3. We find that incorporating EFT fittings can lower PA-MPJPE by more than 20 *mm* across different datasets. This aligns with the observations in [14, 73] that SMPL parameters (θ and β) serve as a more powerful supervision signal than either 2D or 3D keypoints. Cai et al.[73] further explained this by noting that strong supervision enables gradient flow to reach the learnable SMPL parameters via the most direct path.

Remark #1: The indoor/outdoor settings or number of data points are not strong indicators for the model performance. Some attributes (e.g., human pose and shape, camera characteristics, backbone features) are more critical, and having high diversities (leading to considerable overlap between the training and test sets distributions) can give more satisfactory results. Occlusion (person-person or person-object) and SMPL fittings can also help boost recovery accuracy.

3.3.2 Combination of Multiple Datasets

It is a common practice to train the mesh recovery model with multiple datasets of different domains and annotation types. Past works select the datasets empirically. We contend that different combinations of datasets can cause substantial fluctuations in performance. We explore their impacts from two directions.

Selection of datasets. We begin by evaluating different combinations of training datasets, as listed in Table 3.4. Particularly, *Mix 2* follows the configuration used in DSR [16] and EFT [14] while *Mix 6* adopts that of PARE [4]. We have several observations. First, the choice of the training sets has a major impact on the model performance, often exceeding the influence of the training algorithm itself. For example, with *Mix 2*, our HMR baseline achieves a PA-MPJPE of 55.98 mm, whereas DSR and EFT, which employ more advanced methods, report 54.1 mm and 54.7 mm, respectively. The performance gap here is modest compared to

TABLE 3.4: **HMR model performance when trained with different combinations of datasets.**

Mix	Datasets	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
1	H36M, MI, COCO	66.14	115.19	89.04	135.68
2	H36M, MI, EFT-COCO	55.98	91.68	73.17	107.39
3	H36M, MI, EFT-COCO, MPII	56.39	94.56	74.88	111.40
4	H36M, MuCo, EFT-COCO	53.90	87.76	71.10	104.59
5	H36M, MI, COCO, LSP, LSPET, MPII	64.55	109.73	86.62	128.93
6	EFT-[COCO, MPII, LSPET], SPIN-MI, H36M	55.47	90.77	72.78	107.08
7	EFT-[COCO, MPII, LSPET], MuCo, H36M, PROX	52.96	**86.00**	**70.34**	104.49
8	EFT-[COCO, PT, LSPET], MI, H36M	55.97	91.34	73.63	107.90
9	EFT-[COCO, PT, LSPET, OCH], MI, H36M	55.59	89.91	73.20	106.17
10	PROX, MuCo, EFT-[COCO, PT, LSPET, OCH], UP-3D, MTP, Crowdpose	57.80	96.41	75.01	113.55
11	EFT-[COCO, MPII, LSPET], MuCo, H36M	**52.54**	86.68	70.63	**103.07**

the differences observed when altering the dataset mix. Similarly, for *Mix 6*, our baseline obtains 55.47 mm compared to PARE's 52.3 mm. These results indicate that comparisons across models trained on substantially different dataset configurations are potentially unfair. The absence of a standardized dataset mix complicates direct algorithm comparisons. Our benchmarks address this by providing new baselines for several widely used configurations.

Second, adding more datasets does not necessarily yield better performance. As shown in Table 3.4, *Mix 10* underperforms compared to smaller combinations. Excessive dataset diversity can harm accuracy, suggesting that optimal selection matters more than quantity. In our experiments, including EFT datasets—especially EFT-COCO—consistently boosts performance, and should be considered a strong baseline choice (Table 3.4).

While our benchmarking relies on heuristic dataset selection, the process is guided rather than arbitrary. Our analysis suggests that greater overlap between the train–test distributions of key features (e.g., camera parameters, pose, shape, and backbone features) improves performance. This insight allows for selecting the top N datasets to cover a broad attribute space. We also find that certain properties, such as the presence of SMPL annotations, make a dataset more effective for training—yet such datasets remain limited. These findings provide a framework for informed dataset selection, and automating this process will be explored in future work.

Dataset contributions. Beyond the choice of datasets, the relative contribution of each dataset in a mixed set also plays a critical role—yet this aspect has been largely overlooked in prior work. Contributions can be adjusted in two main ways:

TABLE 3.5: **HMR model performance when trained with different contribution configurations of six datasets. (Top) Direct partition. (Bottom) Reweight samples.**

Partition						PA-MPJPE↓
H36M	MI	LSPET	LSP	MPII	COCO	
0.35	0.15	0.10	0.10	0.10	0.20	64.55
0.10	0.10	0.10	0.05	0.15	0.50	61.66
0.20	0.10	0.10	0.05	0.15	0.40	61.23
0.40	0.20	0.10	0.10	0.10	0.10	66.33
0.17	0.17	0.17	0.17	0.17	0.17	63.10

Weighting						PA-MPJPE↓
H36M	MI	LSPET	LSP	MPII	COCO	
0.17	0.17	0.17	0.17	0.17	0.17	63.25
0.10	0.10	0.10	0.05	0.15	0.50	62.43
0.20	0.10	0.10	0.05	0.15	0.40	62.47
0.20	0.10	0.15	0.10	0.15	0.40	63.51
0.35	0.15	0.10	0.10	0.10	0.20	64.93

(1) setting dataset sampling partitions (i.e., the probability of being selected during training) using predefined ratios [2], or (2) keeping partitions fixed while reweighting samples from different datasets, similar to prior importance-weighting strategies [199, 200]. Table 3.5 presents results for various contribution settings using the six datasets from [2]. Our observations are: (1) Adjusting dataset contributions can noticeably affect performance, and carefully increasing the weight of key datasets (e.g., COCO in our case) can bring substantial improvements. (2) When using identical contribution settings, directly changing the sampling partitions is generally more effective than reweighting individual samples.

Remark #2: The selection of datasets and their relative contributions are important factors to determine the model performance. To fairly evaluate and compare the impact of other factors (e.g., training algorithms), it is crucial to keep the same dataset combination configuration, which is usually ignored by prior works. To get a good baseline model, it is suggested to adopt more critical datasets and increase their contribution during training.

3.3.3 Annotation Quality

Many mesh recovery methods rely on pseudo-annotations during training. Here, we examine how annotation quality influences model performance. To this end,

we generate datasets with controlled noise to emulate different levels of corruption observed in real scenarios, and evaluate the following aspects.

Proportion of noisy samples. We inject noise into varying proportions of samples under two scenarios: (1) only the SMPL annotations are corrupted, simulating cases where challenging poses are inaccurately fitted; (2) both keypoints and SMPL annotations are corrupted, reflecting errors from keypoint estimation that propagate to SMPL fittings [14].

Fig. 3.4a presents the performance on the 3DPW test set across different noise ratios. In scenario (1), small amounts of noisy SMPL annotations (<50%) have minimal effect, as the model can still learn effectively from the clean samples. However, when noisy SMPL samples exceed clean ones, errors increase sharply. In scenario (2), adding noisy keypoints on top of noisy SMPL causes large jumps in error, especially beyond 50% corruption. This suggests that with noisy SMPL alone, clean keypoints still offer useful supervision, but when both are noisy, performance degrades severely.

Scale and location of noise. We also investigate: (1) adding Gaussian noise of varying standard deviations to all SMPL pose parameters, following [201, 202], producing still-realistic poses; (2) replacing a proportion of pose parameters for specific body parts (e.g., feet or hands) with random noise, mimicking common fitting errors.

As shown in Fig. 3.4b, increasing Gaussian noise scale slightly raises errors, but they remain low (<70). In contrast, replacing even a fraction of body part parameters with random noise results in large error increases. This indicates that while SMPL annotations should be reasonably accurate, they need not be perfect. Slight and realistic noisy samples can still be beneficial for training, but severe localized corruption is detrimental.

Remark #3: Noisy data samples can harm the model performance, especially when the ratio of samples is higher, or both the SMPL annotation and keypoints are compromised. Slightly noisy SMPL still helps training.

TABLE 3.6: **HMR model performance with different backbone architectures.**

Backbone	Params (M)	FLOPs (G)	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
ResNet-50 [71]	28.79	4.13	64.55	112.34	89.05	130.41
ResNet-101 [71]	47.78	7.83	63.36	112.67	82.65	129.71
ResNet-152 [71]	63.42	11.54	62.13	107.13	81.45	123.95
HRNet-W32 [72]	36.69	11.05	64.27	108.32	82.86	122.36
EfficientNet-B5 [203]	33.62	0.03	65.16	118.15	83.88	144.23
ResNext-101 [204]	91.39	16.45	64.95	114.43	87.26	130.28
Swin [206]	51.72	32.48	62.78	110.42	84.88	137.26
ViT [205]	91.07	11.29	62.81	111.46	84.01	127.22
Twin-SVT [207]	59.27	8.35	60.11	**100.75**	**79.00**	**121.05**
Twin-PCVCT [207]	47.02	6.45	**59.13**	103.85	80.62	123.93

3.4 Benchmarking Backbone Models

Model architecture. Following HMR [2], ResNet-50 [71] is the default backbone in many mesh recovery works [3, 4, 7, 14]. More recently, [4] adopted HRNet-W32 [72] and attributed the performance gains to its ability to produce more robust high-resolution representations. We further consider other architectures. Particularly, we compare different variations of CNN-based models (ResNet-101, ResNet-152, HRNet, EfficientNet [203], ResNext [204]), as well as the latest transformer-based architectures (ViT [205], Swin [206], Twins (-SVT and -PCVCT) [207]).

Table 3.6 reports the performance of the HMR model trained with different backbone architectures. First, increasing the backbone capacity allows deeper feature representations to be learned, yielding performance gains. For instance, the PA-MPJPE is reduced when we switch the backbone model from ResNet-50 to ResNet-152. This is consistent with the findings in [73]. Second, transformer-based backbones are superior to CNN-based backbones, achieving lower PA-MPJPEs and similar FLOPs under comparable parameters (Table 3.6). They are capable of mining rich structured patterns, which are especially essential for learning from different data sources. This contradicts the discoveries in [73], which did not find the advantage of vision transformers over CNN-based ones.

Weight initialization. It is common and computationally efficient to build the HMR model based on a pre-trained backbone. Initialization of the backbone model weights has a significant impact on the HMR model performance. PARE [4] is the first work to use weights from a pose estimation task. It initializes the weights of the HRNet-W32 backbone from a pose estimation model trained on MPII. The initialized model is further finetuned on EFT-COCO for 175K steps

TABLE 3.7: **HMR model performance with different weight initializations.**

Backbone	Mixed Datasets	Dataset for weight initialization		
		ImageNet	**MPII**	**COCO**
ResNet-50	HMR/SPIN	64.55	60.60	57.26
HRNet-W32	HMR/SPIN	64.27	55.93	54.47
Twin-SVT-B	HMR/SPIN	60.11	56.80	52.61
HRNet-W32	PARE	54.84	51.50	49.54

before training on *Mix 6*. Kocabas et al. [4] noted that this strategy accelerates the model convergence and reduces overall training time. However, this study does not provide ablation studies to explore the effect of using pretrained weights from a pose estimation model.

To disclose the impact of weight initialization, we systematically benchmark strategies where the backbone weights are pre-trained with ImageNet, or from pose estimation models trained over MPII or COCO. The results are reported in Table 3.7. First, we observe that transferring knowledge from a strong pose estimation model is sufficient to achieve large improvement gains without having to fine-tune on EFT-COCO, as done in PARE. In Table 3.7, with the HRNet-W32 backbone and weights initialized from MPII, we can already achieve a PA-MPJPE of 51.5 *mm*, which is very close to the error of 50.9 *mm* reported by PARE [4]. The effectiveness of such a pretrained backbone suggests that features learnt from pose estimation tasks are robust and complementary for mesh recovery tasks. Second, the choice of the pose estimation dataset for weight initialization is also vital. Regardless of the backbone variants, pretraining the backbone with COCO gives better performance than MPII for different training dataset mixes and backbone architectures.

Remark #4: The backbone architecture and weight initialization are vital for the HMR performance. Optimal configurations comprise of transformer-based backbones with weights initialized from a strong pose estimation model trained on in-the-wild datasets.

3.5 Benchmarking Training Strategies

3.5.1 Augmentation

Various augmentation strategies have been explored for mesh recovery. SPIN [3] applied rotation, flipping, and color jittering. PARE [4] and BMP [39] incorporated synthetic occlusion by compositing random nonhuman objects into the image, while BMP [60] further made this process keypoint-aware by selectively occluding keypoints. Georgakis et al.[55] controlled occlusion severity by varying the occluder's pattern (oriented bars, circles, rectangles) and size. Mehta et al. [50] introduced inter-person occlusion within the dataset. Beyond occlusion, crop augmentation has been used to improve reconstruction of heavily truncated people [4, 14, 75], and augmentation has also been leveraged to help bridge the synthetic–real domain gap [5, 76–78]. However, these prior works use different configurations and benchmarks, making it difficult to draw general conclusions about augmentation effectiveness.

In this work, we conduct a systematic comparison of nine image-based augmentation techniques across different training and testing setups. We re-implement augmentation methods commonly used in mesh recovery, including random occlusion (hard erasing) [58, 208], synthetic occlusion [4], and crop augmentation [4, 14].

In addition, we incorporate augmentations popular in person re-identification and pose estimation, such as soft erasing [209], self-mixing [209], photometric distortion [210], and coarse/grid dropout [210]. Examples of each augmentation are shown in Fig. 3.5.

We study the effects of data augmentation on two training datasets with distinct characteristics: the indoor H36M set and the outdoor EFT-COCO set. HMR models are trained on each dataset with different augmentations and evaluated on five test sets: 3DPW, EFT-LSPET, EFT-OCHuman, EFT-COCO, and H36M. As shown in Table 3.8, augmentation impacts the two training datasets in different ways. For H36M, almost all augmentations lead to lower errors on outdoor test sets, with self-mixing being the most effective. Without augmentation, the model exhibits an increasing indoor–outdoor domain gap during training, as reflected by rising errors on 3DPW. By contrast, adding augmentation mitigates this gap and reduces overfitting, as evidenced by the training curves in Fig.3.6. Self-mixing is particularly beneficial, producing camera feature distributions that closely match those predicted

TABLE 3.8: **HMR model performance on test sets of 3DPW [65], EFT-LSPET [14], EFT-OCH [14] and H36M [43] and validation set of EFT-COCO [14] when trained on H36M and EFT-COCO with different augmentations. Blue: Augmentation improves the performance. Red: Augmentation harms the performance. Bold: best in column. Underline: second best in column.**

Augmentation	H36M-train					EFT-COCO-train				
	3DPW↓	LSPET↓	OCH↓	COCO↓	H36M↓	3DPW↓	LSPET↓	OCH↓	COCO↓	H36M↓
No augmentation	124.55	207.45	161.77	165.03	53.73	62.37	131.71	**115.50**	114.59	118.39
Hard erasing	107.03	201.16	153.87	147.00	51.70	64.77	136.90	118.93	115.61	120.78
Soft erasing	107.10	193.33	149.93	143.51	47.77	65.70	139.21	118.29	**100.01**	131.09
Self mixing	**101.10**	191.70	**136.68**	**132.17**	**45.12**	63.98	133.18	118.30	125.32	104.37
Photometric distortion	113.53	**190.60**	155.57	153.95	48.45	**62.07**	**128.45**	116.47	112.88	118.92
Random crop	110.08	205.91	150.33	147.27	52.53	71.21	148.80	124.14	104.43	**100.43**
Synthetic occ.	101.96	221.79	146.44	143.32	48.27	63.94	135.00	116.25	103.36	107.14
Synthetic occ. (kp)	107.68	215.34	153.90	145.70	52.26	71.35	142.93	121.34	100.90	103.79
Grid dropout	117.45	208.49	161.69	158.27	57.20	66.65	139.71	118.89	100.52	103.07
Coarse dropout	124.99	202.74	162.50	159.48	50.61	62.78	128.61	116.58	119.70	127.92

by a robust model performing well on 3DPW-test (Fig.3.7a). In contrast, applying augmentation to EFT-COCO has little benefit and sometimes harms performance on 3DPW, EFT-LSPET, and EFT-OCHuman, with improvements observed only on EFT-COCO-val. This aligns with [14], which noted that EFT-COCO-train already contains diverse samples with challenging occlusions and varied camera angles. Consequently, additional augmentation has only a minor effect on the predicted camera feature distribution (Fig. 3.7b) and provides limited gains on out-of-domain benchmarks.

Remark #5: The effect of data augmentations highly depends on the characteristics of the training dataset. Their benefits are more obvious when the training sets contain less diverse and robust samples. When combining multiple datasets during training, we can selectively apply data augmentations to different datasets based on their characteristics.

3.5.2 Training Loss

Prior work in pose estimation often employs the MSE loss for keypoint supervision [211, 212]. In HMR, both keypoints and SMPL parameter regression are typically optimized using MSE. We explore an alternative formulation by replacing MSE with L1 loss. Unlike MSE, the L1 loss measures the magnitude of errors without penalizing their direction, making it less sensitive to outliers. Under certain assumptions, [213] theoretically showed that L1 loss can offer robustness to noisy labels. Motivated by this, we adopt L1 loss for both keypoint and SMPL parameter regression in HMR.

TABLE 3.9: **HMR model performance with and without L1 loss under multi-dataset setting**.

Mix	Datasets	w/oL1	w/L1
1	H36M, MI, COCO	66.14	**57.01**
2	H36M, MI, EFT-COCO	55.98	**55.25**
5	H36M, MI, COCO, LSP, LSPET, MPII	64.55	**58.20**
6	EFT-[COCO, MPII, LSPET], SPIN-MI, H36M	55.47	**53.62**
8	EFT-[COCO, PT, LSPET], MI, H36M	55.97	**53.43**
7	EFT-[COCO, MPII, LSPET], MuCo, H36M, PROX	53.44	**52.93**
11	EFT-[COCO, MPII, LSPET], MuCo, H36M	**52.54**	53.17

This change benefits the model in two main ways. First, it enhances robustness to noisy annotations. Fig. 3.8 compares the performance of HMR trained with MSE versus L1 loss under varying SMPL noise scales and different proportions of noisy keypoints and SMPL labels. The results show that L1 loss consistently yields greater resilience to noise.

Second, L1 loss improves performance in multi-dataset training. Table 3.9 reports results with and without L1 loss for various dataset combinations. Notably, for the dataset configuration used in SPIN [3], applying L1 reduces the PA-MPJPE from 64.55 *mm* to 58.20 *mm*. While the performance boost is most pronounced for suboptimal dataset mixes, the gains diminish when using more optimal selections (*Mix 2, 6, 8*).

Remark #6: Prior works adopt MSE loss for regression of keypoints and SMPL parameters. Using L1 loss instead can not only improve the model's robustness against noisy samples, but also enhance the model performance, especially when the selected datasets are not optimal.

3.6 Benchmarking Other Algorithms and Test Sets

In the previous benchmarking experiments, we used the HMR algorithm and the 3DPW test set. However, our evaluation methodology and conclusions are applicable to other algorithms and benchmarks as well. In this section, we present additional experiments to validate their generalization.

Other algorithms. Beyond HMR, we evaluate several widely used approaches, including SPIN [3], GraphCMR [10], PARE [4], and Graphormer [62]. Table 3.10

TABLE 3.10: **Model performance (3DPW-test PA-MPJPE in *mm*) when trained with different recommended strategies of L1 loss, weight initialisation from COCO pose estimation model, and selective augmentation.**

Algorithms	Datasets	Backbones	Initialisation	Normal	L1	L1+COCO	L1+COCO+Aug
HMR	H36M, MI, COCO, MPII, LSP, LSPET	ResNet-50	ImageNet	64.55	58.20	51.80	51.66
SPIN	H36M, MI, COCO, MPII, LSP, LSPET	ResNet-50	HMR (ImageNet)	59.00	57.08	51.54	50.69
GraphCMR	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	ImageNet	70.51	67.20	61.74	60.26
PARE	EFT-[COCO, LSPET, MPII], H36, MI	HRNet-W32	ImageNet	61.99	61.13	59.98	58.32
Graphormer	H36M, COCO, UP3D, MPII, MuCo	HRNet-W48	ImageNet	63.18	63.47	59.66	58.82

TABLE 3.11: **Model performance of other algorithms with optimized configurations on the 3DPW test set.** Abbreviations for the datasets - Human3.6M [43]: H36M, MPI-INF-3DHP [44]: MI, MuCo-3DHP [50]: MuCo

Algorithm	Dataset	Backbone	Variant	PA-MPJPE↓	MPJPE↓	PA-PVE↓	PVE↓
PARE [4]	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	EFT-COCO	50.90	82.0	-	97.9
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	-	61.99	109.82	82.33	133.86
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HrNet-W32	L1-COCO-Aug	58.32	100.35	77.22	121.97
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twins-SVT	L1-COCO-Aug	51.96	93.46	81.33	130.20
PARE (Ours)	EFT-[COCO, LSPET, MPII], H36M, MuCo	Twins-SVT	L1-COCO-Aug	51.93	91.43	68.40	110.32
GraphCMR [10]	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	-	70.52	116.83	87.50	133.67
GraphCMR	COCO, H36M, MPII, LSPET, LSP, UP3D	ResNet-50	L1-COCO-Aug	60.26	99.28	75.75	113.17
GraphCMR	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	ResNet-50	-	60.51	101.69	77.51	121.37
GraphCMR	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	Twins-SVT	L1-COCO-Aug	53.29	91.07	70.52	108.14
SPIN [3]	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	-	59.2	96.9	-	135.1
SPIN (Ours)	H36M, MI, COCO, LSP, LSPET, MPII	ResNet-50	L1-COCO-Aug	50.54	80.49	68.29	96.67
SPIN (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	ResNet-50	-	55.28	93.52	72.19	109.57
SPIN (Ours)	EFT-[COCO, LSPET, MPII], H36M, SPIN-MI	HRNet-W32	L1-COCO-Aug	47.59	80.77	64.22	96.22
MeshGraphormer [62]	H36M, COCO-2017, UP3D, MPII, MuCo	HRNet-W48	-	63.18	108.02	76.05	125.56
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, MPII, MuCo	HRNet-W48	L1-COCO-Aug	58.82	104.63	76.79	132.52
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, MPII, MuCo	Twins-SVT	L1-COCO-Aug	58.13	98.03	73.32	116.95
MeshGraphormer (Ours)	H36M, COCO-2017, UP3D, EFT-MPII, MuCo	Twins-SVT	L1-COCO-Aug	58.30	96.71	74.88	124.97

reports the performance of these algorithms under various configurations[2]. Table 3.11 extends this comparison to different dataset mixes and backbones. Similar to HMR, we find that high-performing models for other algorithms are obtained when using L1 loss, weight initialisation from COCO pose estimation model, and selective augmentation.

Other test sets. While 3DPW is the most common evaluation benchmark, prior work has also tested on MuPoTs-3D-test [39], AGORA-test [214], MPI-INF-3DHP-test [58], and challenging benchmarks such as EFT-OCHuman-test and EFT-LSPET-test [14]. For completeness, we evaluate on seven additional benchmarks: H36M, AGORA validation, MPI-INF-3DHP test, EFT-COCO validation, MuPoTs-3D test, EFT-OCHuman test, and EFT-LSPET test. We perform dataset benchmarking across all selected test sets and compute the correlation between model performance

[2]Our baseline models for HMR, SPIN and GraphCMR can reach the reported results in the respective works. For PARE, the original work trains the model on MPII for pose estimation task and later on EFT-COCO for mesh recovery before training on the full set of datasets. To keep consistent with the practice adopted throughout our work, we benchmark PARE by training it from scratch with only ImageNet initialisation. For Graphormer, the original work evaluates on H36M every epoch before fine-tuning the best H36M model on 3DPW-train (Protocol 1) for 5 epochs. To keep consistent with the evaluation settings throughout this work, we train each model for 100 epochs and report the best PA-MPJPE on 3DPW-test set (Protocol 2). We provide the training logs for all the experiments in https://github.com/smplbody/hmr-benchmarks.

TABLE 3.12: **Correlation of performance on test benchmarks**

	EFT-COCO	3DPW	AGORA	EFT-OCH	EFT-LSPET	MI	MuPots-3D	H36M	Average
EFT-COCO	1.000	0.860	0.891	0.910	0.820	0.643	0.595	0.387	0.729
3DPW	0.860	1.000	0.768	0.761	0.779	0.704	0.396	0.506	0.682
AGORA	0.891	0.768	1.000	0.793	0.624	0.626	0.696	0.183	0.654
EFT-OCH	0.910	0.761	0.793	1.000	0.750	0.449	0.424	0.378	0.638
EFT-LSPET	0.820	0.779	0.624	0.750	1.000	0.562	0.372	0.438	0.621
MI	0.643	0.704	0.626	0.449	0.562	1.000	0.640	0.246	0.553
MuPots-3D	0.595	0.396	0.696	0.424	0.372	0.640	1.000	0.104	0.461
H36M	0.387	0.506	0.183	0.378	0.438	0.246	0.104	1.000	0.320

on each. As shown in Table 3.12, performance on 3DPW correlates well with results on other benchmarks, supporting its suitability as a representative evaluation set. This contrasts with H36M, where strong results do not reliably indicate performance on other test sets.

3.7 Lessons from Our Benchmarking

We summarise our findings and open questions raised by these findings:

Datasets.

1. The choice of datasets and their relative contributions are key determinants of model performance. To fairly evaluate other factors (e.g., training algorithms), the dataset combination must be kept consistent—something often overlooked in prior works.

2. Diversity of attributes (e.g., human pose and shape, camera characteristics, backbone features) in training datasets is critical. High diversity, which increases the overlap with test-set distributions, generally leads to better performance. *Leveraging knowledge of train–test attribute distributions, merging datasets that collectively span a broad range of attributes can be highly effective. This diversity can guide future dataset creation, enhancement, and selection.*

3. Adjusting dataset contributions by directly altering partitions (thus increasing the share of valuable samples) is more effective than keeping partitions fixed and reweighting samples. To obtain strong baselines, it is advisable to prioritize critical datasets and increase their contributions during training. However, current partitioning is still manual. *An open question is how to automate dataset selection and contribution adjustment; one possible direction is to treat partitions as hyperparameters and tune them using AutoML.*

4. Adding SMPL fittings—even if slightly noisy—remains highly beneficial for training, whereas noisy keypoints significantly harm performance. *Adding pseudo-annotations to existing outdoor 2D-keypoint datasets could be a cost-effective way to improve them.*

5. Robust test sets should: (1) contain accurate ground-truth SMPL annotations captured via mocap or simulation. While EFT-COCO-Val appears representative, our visualizations reveal annotation errors, highlighting the limitations of pseudo-annotated datasets for benchmarking. Currently, 3DPW is the only large-scale real-world dataset with reliable SMPL ground truth; and (2) exhibit sufficient diversity to reflect real-world conditions. In contrast, H36M is not strongly indicative of generalization, and relying on it as a primary benchmark risks overestimating performance. *These observations suggest the need for an additional in-the-wild test set beyond 3DPW. Such a dataset should combine reliable ground-truth annotations with complementary scene diversity (e.g., indoor environments, occlusions, varied body shapes) and sufficient scale. Without such a benchmark, performance on 3DPW alone may not reliably reflect generalization in real-world settings.*

Backbone and initialization

1. Fair algorithm evaluation requires proper ablations on backbones and initialization using standard baselines.

2. Optimal configurations use transformer-based backbones initialized from strong pose estimation models trained on in-the-wild datasets. *This highlights the value of transferring knowledge from pose estimation to mesh recovery, motivating careful consideration of dataset usage in both tasks.*

Training strategies

1. The benefit of data augmentation depends heavily on the diversity of the training dataset. Augmentation is most effective when the dataset lacks diversity or robustness. In multi-dataset training, augmentation can be applied selectively to datasets based on their characteristics. *Augmentation can enhance indoor datasets that, while accurate, have limited diversity.*

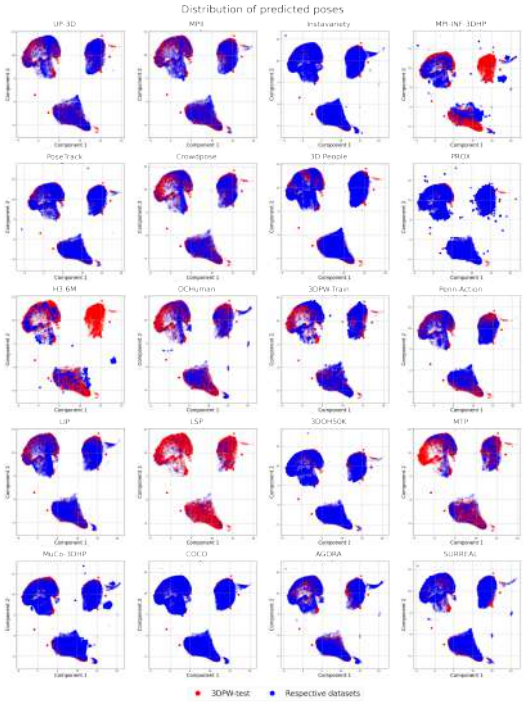
2. While MSE loss is commonly used for regressing keypoints and SMPL parameters, replacing it with L1 loss improves robustness to noisy samples and can boost performance—especially when dataset selection is suboptimal.

3.8 Conclusion

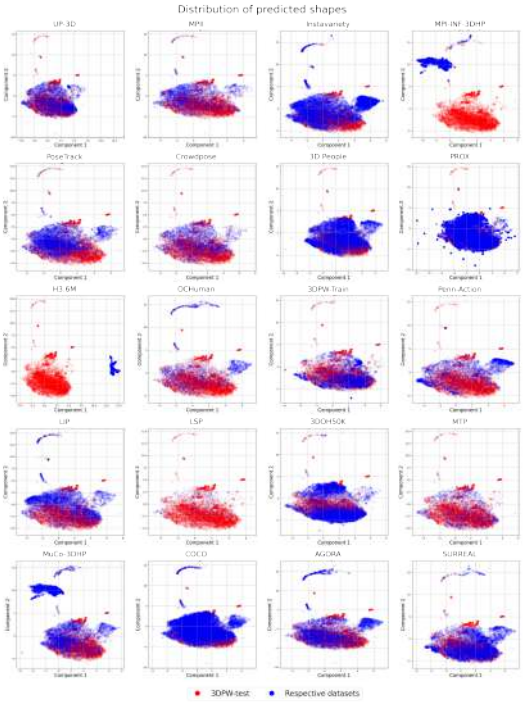
Large amounts of efforts have been devoted to the exploration of novel algorithms for 3D human mesh recovery. However, there are also other important factors that can affect the model performance, which are rarely investigated in a systematic way. To the best of our knowledge, this work presents the *first* large-scale benchmarking of various configurations for mesh recovery tasks. We identify the key strategies and remarks that can significantly enhance the model performance. We believe this benchmarking study can provide strong baselines for unbiased comparisons in mesh recovery studies. We summarize all our findings in Section 3.7.

Future works. There are a couple of future research directions. (1) Due to the large amount of experiments, we mainly perform the benchmarks on HMR, which is an important milestone work with straightforward architecture. We provide some evaluation results on a few other algorithms in Section 3.6 to show the generalization of our major findings. In the future, we plan to extend our studies to more 3D human pose and mesh reconstruction algorithms. (2) Currently we need to use prior knowledge to manually select the datasets and their partitions. Future efforts could investigate if it would be possible to automatically determine the optimal selection of datasets and partitions. For instance, we find that dataset-level weighting is more effective than sample-level weighting. If we consider dataset partition as a hyperparameter to tune, we can borrow techniques from automatic hyperparameter tuning with methods such as reinforcement learning or bayesian optimization to automate dataset configurations. (3) In this chapter, we experimentally disclose some inspiring conclusions about HMR training. It is worth conducting deeper investigations to interpret and explain those findings, and obtain the optimal strategy. This will be our future work as well.

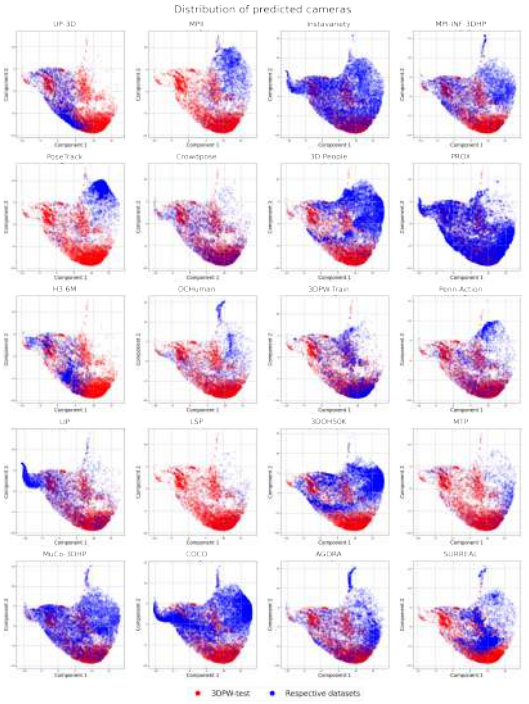
Within the broader thesis narrative (Fig. 1.1), this chapter establishes the factors that drive mesh recovery performance under realistic conditions, motivating the subsequent focus on robustness in Chapter 4.



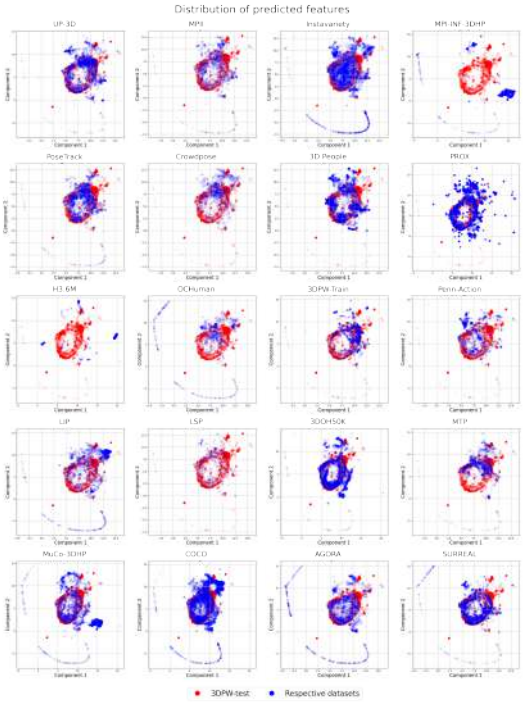
(a) **Poses.**



(b) **Shapes.**



(c) **Estimated cameras.**



(d) **Backbone features.**

FIGURE 3.3: [
Feature distributions of pose, shape, camera, and backbone for 3DPW-test vs other datasets]**Feature distributions of (a) poses, (b) shapes, (c) estimated cameras, and (d) backbone features between 3DPW-test (red) and the respective datasets (blue).**

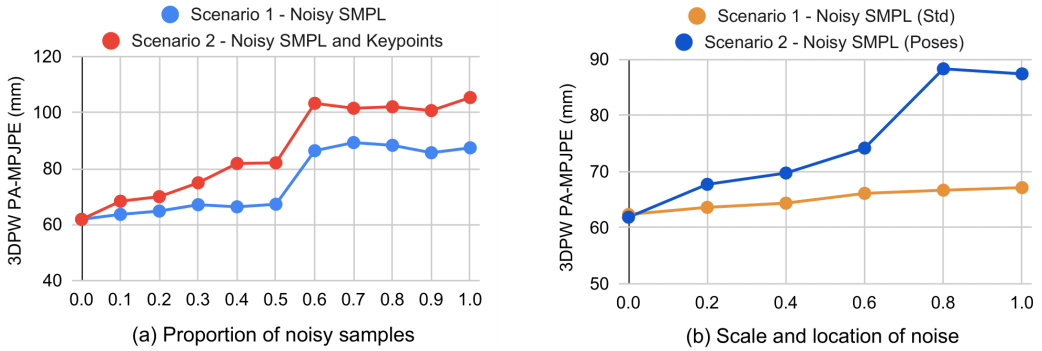


FIGURE 3.4: **HMR model performance with different types of noisy training data.**

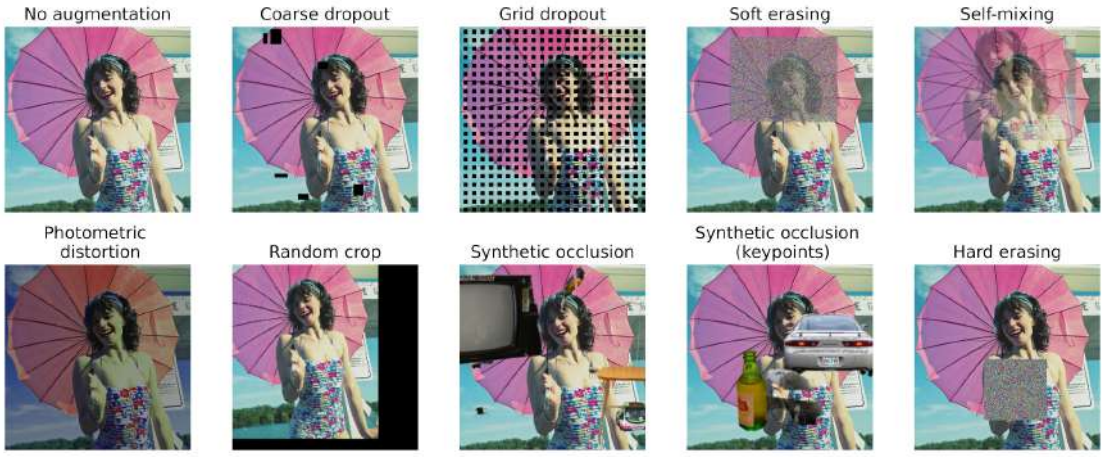


FIGURE 3.5: **Visualisation of augmented samples.**

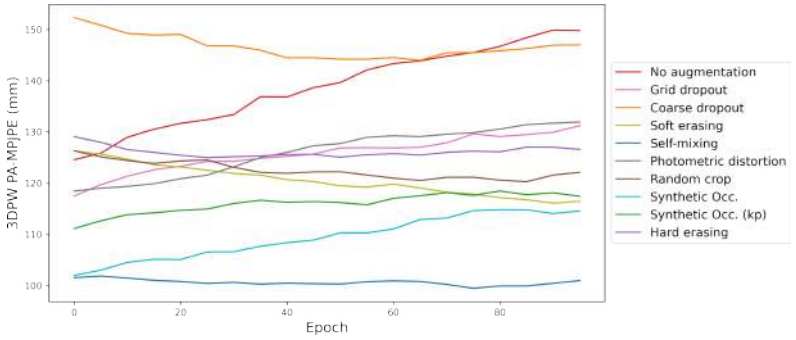


FIGURE 3.6: **Per-epoch evaluation on 3DPW (PA-MPJPE in *mm* when trained on H36M under different augmentations.**

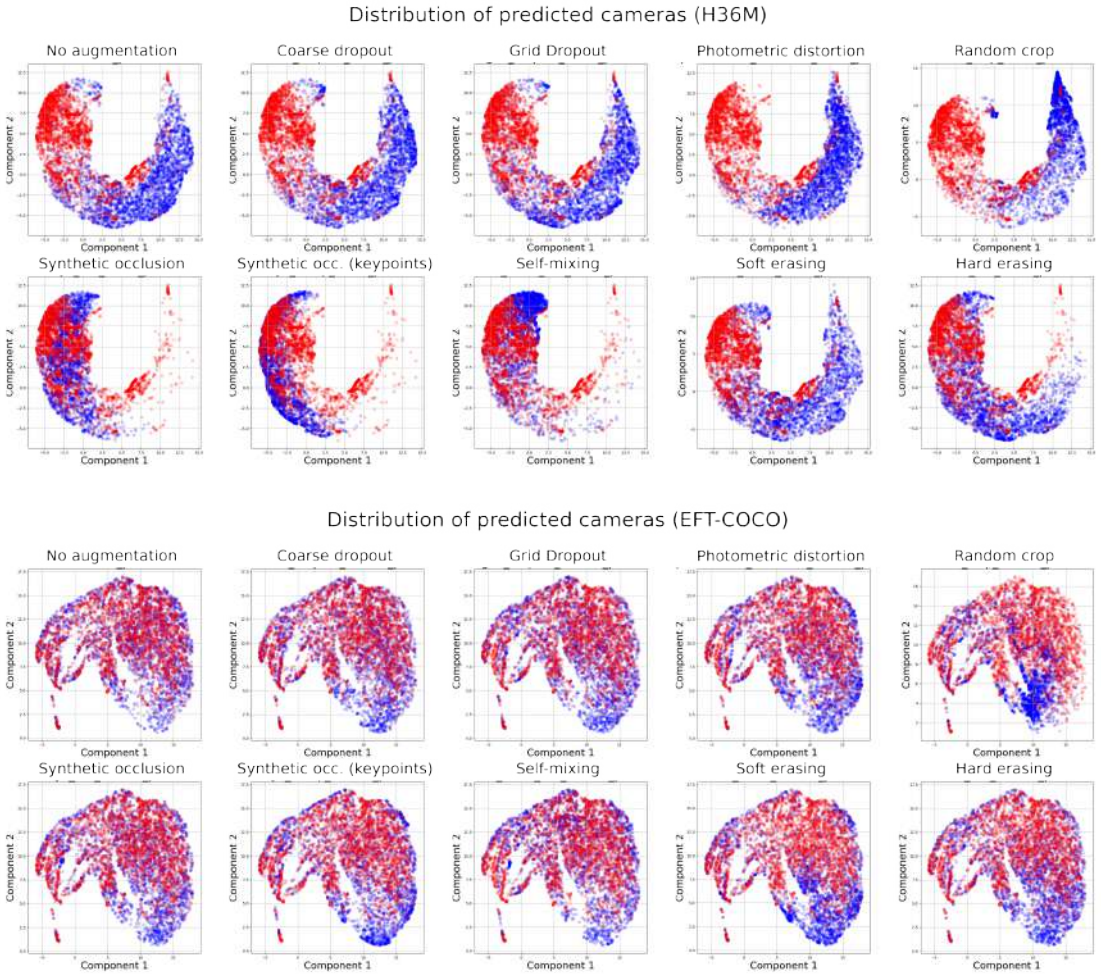


FIGURE 3.7: **Effect of applying augmentation on the distribution of predicted camera features for (top) H36M and (bottom) EFT-COCO.**

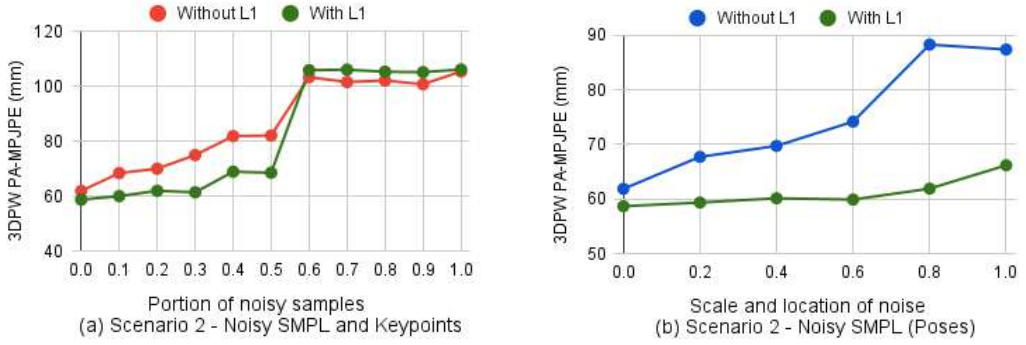


FIGURE 3.8: **HMR model performance with and without L1 loss under different (a) proportions of noisy SMPL and keypoints; (b) ratios of noisy pose parameters.**

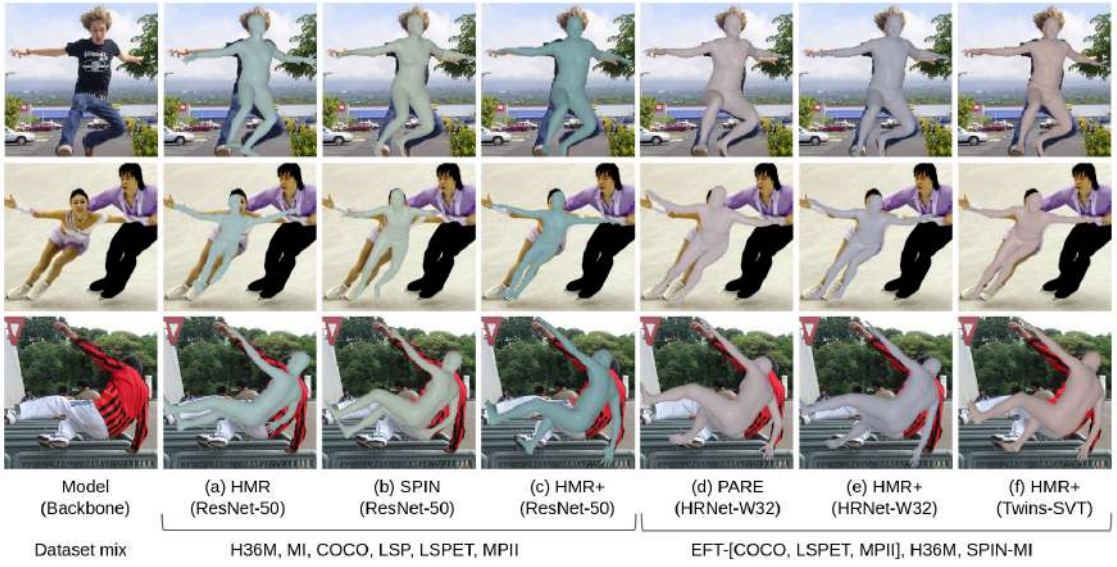


FIGURE 3.9: **Qualitative results on COCO and LSPET test sets. From left to right: (a) HMR [2], (b) SPIN [3], (c) HMR+ (ResNet-50) (d) PARE [4] (e) HMR+ (HRNet-W32) (f) HMR+ (Twins-SVT). (a)-(c) follow [6]'s dataset mix while (d)-(f) follow [4]'s dataset mix. HMR+ adopts COCO-weight initialization, L1 loss and selective augmentation. More examples in [181].**

Chapter 4

Towards Robust and Expressive 3D Human Pose and Shape Estimation

4.1 Introduction

Building on the findings of Chapter 3, which revealed the sensitivity of mesh recovery performance to dataset selection, backbone choice and training strategy, this chapter focuses on improving the robustness of whole-body pose and shape estimation under realistic conditions. We develop a pixel-aligned whole-body SMPL-X estimator that is resilient to imperfect crops and misalignment, providing a stronger body-level foundation for clothed and disentangled avatar modeling in Chapter 5.

Early studies employed parametric statistical models such as SMPL [1], MANO [41], and FLAME [40] to reconstruct different body parts separately, including the human body [2–4, 12, 14–17], face [80, 216, 217], and hands [59, 79, 89]. More recently, interest has shifted toward *whole-body estimation* [57, 85–87, 218], which aims to jointly predict body pose, hand articulation, and facial expressions for a complete 3D representation. Typically, these approaches first process body, hand, and face regions using separate sub-networks to extract part-specific features. The resulting features are then combined to regress whole-body parameters—such as joint rotations, shape coefficients, and expression codes—which are subsequently fused to generate a unified whole-body mesh. This development is a key step toward efficient and practical modeling of human behavior.

The work in this chapter has been published in [215].

Achieving robust and accurate whole-body reconstruction is challenging because it requires precise estimation for each part while maintaining correct connectivity between them. Hands and faces, due to their smaller image sizes, are often localized, cropped, and upsampled before being passed to dedicated sub-networks. In real-world scenarios, where ground-truth bounding boxes are unavailable, current methods rely on detection algorithms to produce these crops. The performance of the entire pipeline is therefore highly dependent on the accuracy of the detected crops. As shown in our experiments in Section 4.2, even small variations in crop scale or alignment can noticeably degrade performance, suggesting a limited capacity of existing models to robustly localize and extract discriminative features from the subject.

Our analysis identifies three critical limitations in current whole-body pose and shape estimation systems: (1) insufficient accuracy in localizing the subject and its individual parts, (2) suboptimal extraction of useful features, and (3) imprecise pixel-level alignment between predicted meshes and input images. To address these, we introduce three dedicated components, each targeting one of these issues:

- **Localization Module**. This module employs both sparse and dense prediction branches to enhance the network's awareness of the spatial position and semantics of each body part in the image. The learned location of the joint positions are helpful in recovering the relative rotations.
- **Contrastive Feature Extraction Module**. This module incorporates a pose- and shape-aware contrastive loss, along with positive samples, to promote better feature extraction under robust augmentations. By minimizing the contrastive loss, the model can produce consistent representations for the same subject, even when presented with different augmentations, making it robust to various transformations and capable of extracting meaningful invariant features.
- **Pixel Alignment Module**. This module utilizes differentiable rendering to enforce precise pixel alignment of the projected mesh and input image, enabling the learning of more accurate pose, shape and camera parameters.

By integrating these modules, our framework, `RoboSMPLX`, delivers improved robustness and accuracy for whole-body 3D mesh recovery across body, face, hand, and combined benchmarks.

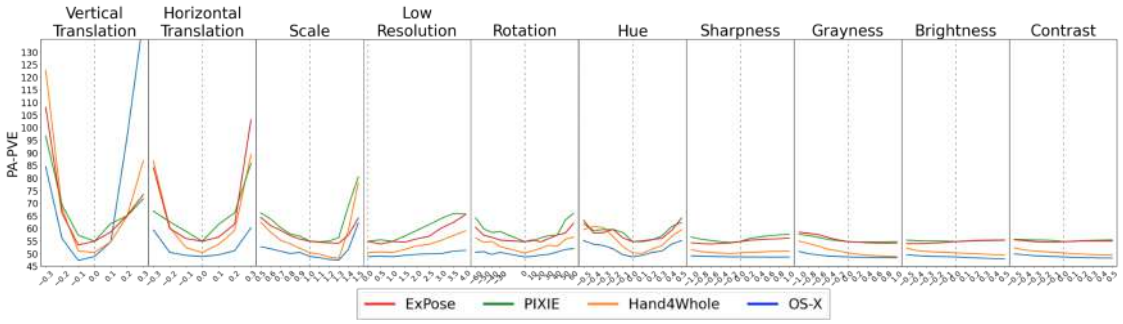


FIGURE 4.1: **Wholebody PA-PVE errors under different augmentations (sorted in descending order).** The dashed line indicates baseline performance without augmentation.

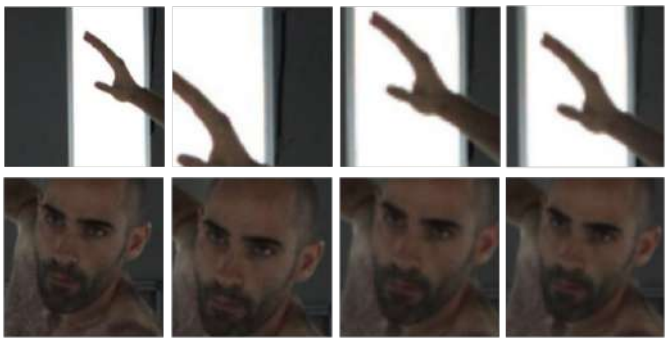


FIGURE 4.2: **Crops from (a) ExPose [85] (b) PIXIE [87], (c) Hand4Whole [13] (d) `RoboSMPLX`.**

4.2 Motivation

As discussed in Section 4.1, current whole-body pose and shape estimation approaches are prone to robustness issues, largely due to their sensitivity to the quality of input crops. To better understand the underlying causes, we perform a systematic evaluation of four state-of-the-art approaches: ExPose [85], PIXIE [87], Hand4Whole [13] and OS-X [84]. We design ten realistic augmentations (details in Appendix of [215]) and vary their intensities within plausible ranges. These augmentations fall into three categories: (1) *image-variant* — modify image appearance without altering the subject's 3D pose or position (e.g., color jittering); (2) *location-variant* — change the subject's location in the frame without altering its pose (e.g., translation, scaling); (3) *pose-variant* — simultaneously modify both the pose and location (e.g., rotation).

Impact of subject localization. We first reveal that existing models demonstrate high sensitivity to the subject's position, indicating potential difficulties in subject

FIGURE 4.3: **Sensitivity of existing body and hand models to different alignments (left) and scales (right).**

localization. Figure 4.1 reports the PA-PVE errors of the whole body under different augmentations. We observe that image-variant augmentations (contrast, sharpness, brightness, hue and grayscale) lead to an acceptable range of error rates (approximately in the 50s) and minimal fluctuation (around ±2). In contrast, location-variant augmentations altering the subject's position within the frame, such as rotation, scaling, and horizontal or vertical translation, cause significantly larger errors, revealing a strong sensitivity to positional changes.

Such transformations are common in real-world settings, where crops are obtained from external detectors and precise control is difficult. To ensure visibility, crops are often expanded; however, smaller scale factors ($<$1.0) notably degrade performance (Figure 4.1). Horizontal or vertical offsets, corresponding to imperfect centering or partial visibility, also harm accuracy. This effect is not limited to whole-body meshes: the scale and alignment of crops similarly affect body-, face-, and hand-specific estimators (Figures 4.3 and 4.8). For whole-body systems, inaccurate part crops (Figure 4.2) degrade the relevant sub-network outputs, propagating errors to the final prediction.

Impact of feature extraction. Performance drops under translation or scale changes, even when the subject remains visible, indicate that current models struggle to extract relevant features related to the subject of interest and effectively disregard irrelevant background elements. This suggests a lack of invariance in the learned features. To enhance the model's robustness, it is critical to produce consistent features irrespective of various augmentations applied to the image.

Impact of output pixel alignment. Pixel alignment is a critical aspect of high model performance. In certain instances, even with accurate subject localization, the model fails to produce properly aligned results (Figure 4.10). This is commonly due to suboptimal camera parameter estimation. High-quality mesh recovery therefore depends on reliably estimating camera parameters so that the projected mesh aligns

precisely with the image. Such alignment directly benefits the accuracy of pose, shape, and camera parameter predictions, and thus the reliability of the entire estimation process.

4.3 `RoboSMPLX` Framework

We design `RoboSMPLX` to strengthen the robustness of whole-body pose and shape estimation. It integrates three targeted modules, each designed to address the challenges outlined in Section 4.2: 1) **Localization Module** (Section 4.3.2): explicitly learns the subject's spatial location and embeds this information into the estimation of pose, shape, and camera parameters ; 2) **Contrastive Feature Extraction Module** (Section 4.3.3): facilitates the extraction of stable and relevant features under diverse augmentations, enhancing generalization and robustness to real-world scenarios; 3) **Pixel Alignment Module** (Section 4.3.4): enforces pixel-level alignment of predictions and image.

We first describe the overall `RoboSMPLX` architecture, which comprises separate Body, Hand, and Face subnetworks (Section 4.3.1). Each subnetwork incorporates the **Localization Module** and **Pixel Alignment Module**, and employs the **Contrastive Feature Extraction Module** to learn more robust features. Figure 4.6 illustrates the Hand subnetwork architecture; the Body and Face subnetworks follow the same design.

4.3.1 Architecture and Training Details

Figure 4.4 shows the overall pipeline of `RoboSMPLX` for whole-body 3D human pose and mesh estimation. The Body subnetwork outputs 3D body joint rotations $\theta_b \in \mathbb{R}^{21\times3}$, global orientation $\theta_{bg} \in \mathbb{R}^3$, shape parameters $\beta_b \in \mathbb{R}^{10}$, camera parameters $\pi_b \in \mathbb{R}^3$, and whole-body joints $K \in \mathbb{R}^{137\times3}$. Joints corresponding to the hand and face are used to derive bounding boxes. Subsequently, hand and face images are cropped from a high-resolution image to preserve details. The Hand subnetwork predicts left and right hand 3D finger rotations $\theta_h \in \mathbb{R}^{15\times3}$. Simultaneously, the Face subnetwork generates 3D jaw rotation $\theta_f \in \mathbb{R}^3$ and expression $\psi_f \in \mathbb{R}^{10}$. When training Hand and Face subnetworks with part-specific datasets, additional

parameters such as global orientation $\theta_{fg} \in \mathbb{R}^3$, shape $\beta_f \in \mathbb{R}^{50}$, and camera $\pi_f \in \mathbb{R}^3$ are estimated. These branches are discarded during whole-body estimation and training.

Body subnetwork. The body image is downsampled from the original image to reduce the computational cost, resulting in $I_b \in R^{3\times256\times256}$. The Body subnetwork outputs 3D body joint rotations $\theta_b \in R^{21\times3}$, global orientation $\theta_{bg} \in R^3$, shape parameters $\beta_b \in R^{10}$, camera parameters $\pi_b \in R^3$, and whole-body joints $K \in R^{137\times3}$. Hand and face bounding boxes are then derived from the face and hand keypoints. Width and height are determined from the x-y range of the keypoints, and the center is the aggregated mean of the keypoints. High resolution crops are used for hand and face inputs following ExPose and PIXIE. In line with ExPose [85] and PIXIE [87], hand and face input images are obtained from high resolution crops to utilize the information available from the original image instead of the downsampled image.

Hand subnetwork. After obtaining the cropped hand images $I_h \in R^{3\times256\times256}$, the left hand images are flipped to match the orientation of the right hands before being input to the Hand subnetwork. After predicting the 3D finger rotations $\theta_h \in R^{15\times3}$, the outputs of the flipped left hands are reverted to their original orientation. The 3D finger rotations of the left and right hands are denoted as θ_{rh} and θ_{lh} respectively. When training the full version on hand datasets, we also output the global orientation $\theta_{hg} \in R^3$, shape $\beta_h \in R^{10}$ and camera $\pi_h \in R^3$. However, these branches are discarded during whole-body estimation and training.

Face subnetwork. This subnetwork generates the 3D jaw rotation $\theta_f \in R^3$ and expression $\psi_f \in R^{10}$ from the cropped face image $I_f \in R^{3\times256\times256}$. When training the full version on face datasets, additional outputs include the global orientation $\theta_{fg} \in R^3$, shape $\beta_f \in R^{50}$, expression $\psi_f \in R^{50}$ and camera $\pi_f \in R^3$. These branches are also discarded during whole-body estimation and training.

Implementation details. We implement our framework with PyTorch [219] and Pytorch3D [220]. For model initialization, we pre-train the ResNet backbone on the MSCOCO 2D whole-body human pose dataset. During training, we use the Adam optimizer with a mini-batch size of 32 and apply data augmentations, e.g., scaling, rotation, random horizontal flip, and color jittering. The initial learning rate is set to 10e−4, decayed by a factor of 10 at the later epoch. We use the SMPL,

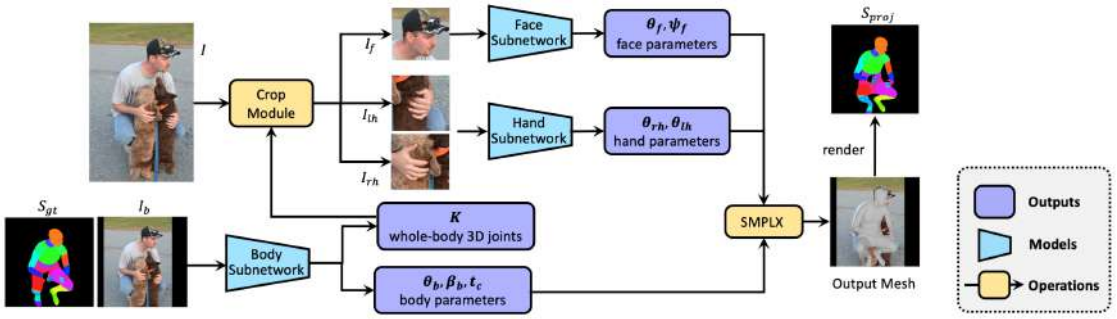


FIGURE 4.4: **Pipeline of our RoboSMPLX framework consisting of Body, Hand and Face subnetworks.**

MANO, FLAME and SMPL-X body models for the training of body, hand, face and wholebody respectively. Further details will be provided in our code.

Subnetworks are trained separately, then integrated in a multi-stage manner. Initial whole-body training runs for 20 epochs. The hand and face modules are substituted with the trained Hand and Face subnetworks, followed by 20 epochs of fine-tuning to better unify the knowledge from the Hand and Face subnetworks into the whole-body understanding. Each subnetwork is trained by minimizing the following loss function L:

$$L = \lambda_{3D}L_{3D} + \lambda_{2D}L_{2D} + \lambda_{BM}L_{BM} + \lambda_{proj}L_{proj} + \lambda_{segm}L_{segm} + \lambda_{con}L_{con} \quad (4.1)$$

Here L_{BM} is the L1 distance between the predicted and ground-truth body model parameters. L_{3D} denotes the L1 distance between 3D keypoints and joints regressed from the body model. L_{2D} signifies the L1 distance of the ground-truth 2D keypoints to predicted and projected 2D joints. The latter are obtained by projecting the regressed 3D coordinates from the 3D mesh to the image space using the perspective projection [2]. The part segmentation loss L_{segm} is the cross-entropy loss between $P_{h,w}$ after softmax and $P_{h,w}$ averaged over H×W elements, following [4]. L_{proj} refers to the projected segmentation loss, which is the sigmoid loss between the projected mesh and the ground-truth segmentation map. L_{con} is the contrastive loss described in Section 4.3.3. For wholebody training, L_{box} is added to measure the L1 distance between the predicted and actual center and scale of the hands' and face's boxes.

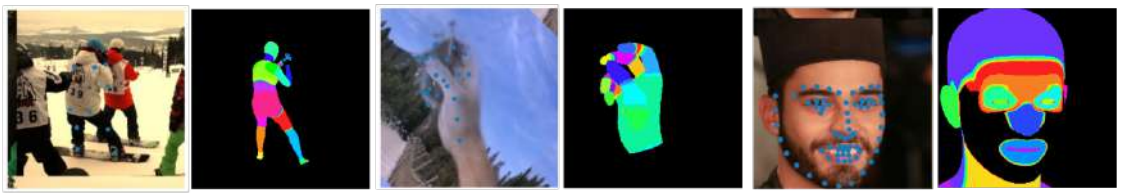


FIGURE 4.5: **Examples of keypoint and part segmentation supervision for Body, Hand and Face subnetworks.**

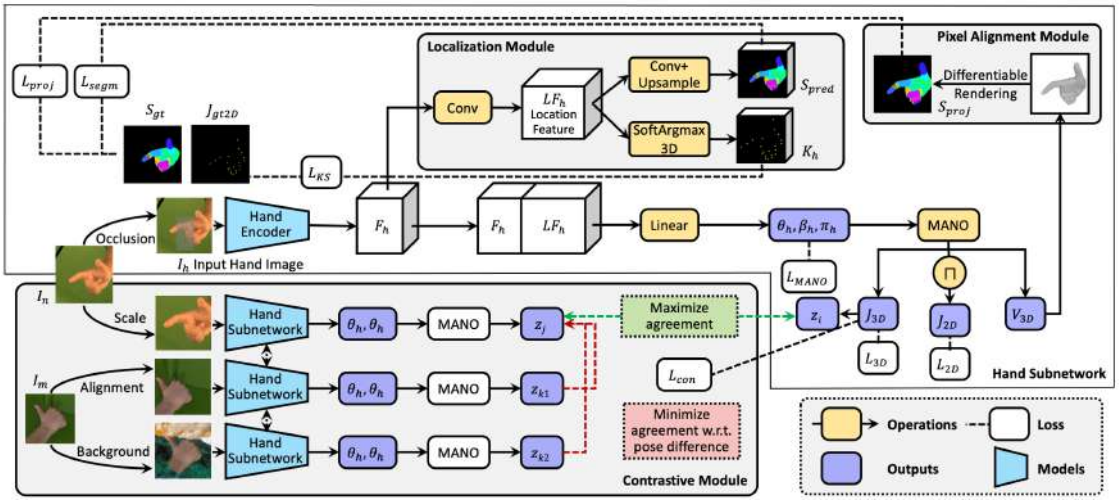


FIGURE 4.6: **Subnetwork Architecture with three modules.** We use the Hand subnetwork as an example. z represents normalized J_{3D} while p corresponds to the ground-truth of z. Green and red dashed lines refers to contrastive loss for positive and negative samples respectively.

4.3.2 Localization Module

This module targets subject localization by learning explicit sparse and dense predictions of the subject within the image. Figure 4.5 shows an example of the supervision used for each subnetwork. Unlike prior approaches that regress pose rotations directly from backbone features, this module injects explicit awareness of the subject's location and semantics into the prediction of pose, shape, and camera parameters. In doing so, it reduces sensitivity to positional variations caused by small shifts in crop scale or bounding-box alignment.

As shown in Figure 4.6, given an image, a convolutional backbone is utilized to extract its feature map $F \in \mathbb{R}^{512\times32\times32}$. Following Hand4Whole [13], a 1×1 convolutional layer is then used to predict 3D feature maps $LF \in \mathbb{R}^{32J\times32\times32}$ from F, where J represents the number of predicted joints with a feature map depth of 32. LF contains valuable information about the mesh's position in the image and

FIGURE 4.7: **Augmentations for the Body subnetwork.** Black, blue and red labels represent image-variant, location-variant and pose-variant augmentations, respectively.

semantics of various parts. It is concatenated with the backbone feature map F to predict pose $\theta \in \mathbb{R}^P$, shape $\beta \in \mathbb{R}^{10}$ and camera translation $\pi \in \mathbb{R}^3$, where P is the number of body parts. Meanwhile, LF is also used to obtain extra information with two branches: (1) 3D joint coordinates $K \in \mathbb{R}^{J\times 3}$ are obtained from LF using the soft-argmax operation [221] in a differentiable manner. (2) 2D part segmentation maps $S \in \mathbb{R}^{P+1\times 64\times 64}$ are extracted from LF with several convolution layers, which model P part segmentation and 1 background mask. Here, 64 represents the height and width of the feature volume, and each pixel (h, w) stores the likelihood of belonging to a body part P.

Learning part segmentation maps and 3D joint coordinates offers complementary benefits. The 3D joint coordinates provide depth cues that can help establish the relative ordering of parts in the segmentation maps, while joints often lie along the boundaries between segments, naturally acting as markers to separate adjacent parts. In our design, the Body subnetwork uses 24 parts P and 137 joints J; the Hand subnetwork handles 16 parts P and 21 joints J; and the Face subnetwork covers 15 parts P and 73 joints J.

4.3.3 Contrastive Feature Extraction Module

This module integrates a pose- and shape-aware contrastive loss together with the use of positive sample pairs. By minimizing this objective, the network learns to generate stable and consistent representations for the same individual under different augmentations, thus enabling more reliable feature extraction and improved robustness in varied conditions.

Conventional contrastive learning methods based on SSL (e.g., SimCLR) face challenges in unifying similar pose embeddings and distancing dissimilar ones in human pose and shape estimation tasks. Without labels for guidance, images

with similar poses could be misidentified as negative samples and contrasted away, complicating the self-organization of the embeddings in pose space. Figures 4.11 show their ineffectiveness for the 3D human pose and shape task [90] by visualizing the retrieved samples from the embeddings. The supervised contrastive learning approach by Khosla et al. [91], though effective for image classification, might not extend well to human pose and shape estimation, which is a high-dimensional regression problem and poses exist in a continuous space rather than well-defined classes.

Our module addresses these challenges with two main contributions. First, we evaluate three pose representations $\boldsymbol{z}$ and their respective distance measures: (1) a concatenated vector of global orientation and rotational pose; (2) global orientation and rotational pose treated as separate components; and (3) 3D root-aligned joints, regressed from the body model and derived from both pose and shape parameters. For (1) and (2), we explore both 6D vector and rotation matrix representations. For (3), we compare L1, Smooth L1, and Mean Squared Error (MSE) distance functions (Table 4.10).

Second, we experiment with ten augmentation types, organized into three categories (Figure 4.7): (1) *image-variant* augmentations, such as color jitter, blur, occlusion, and background replacement; (2) *location-variant* augmentations, including translation and scaling; and (3) *pose-variant* augmentations, such as rotation and horizontal flipping. Our ablations (Table 4.11) show that augmentations altering global orientation degrade performance, so we exclude them when forming positive pairs. Instead, positives are generated by applying a random combination of location-variant and image-variant transformations.

Formally, given a batch of N samples, we generate another N images by applying augmentation to each sample. For an anchor i and its positive j, i is contrasted with $2N-1$ examples (1 positive and $2N-2$ negatives). The loss is defined as:

$$\mathcal{L}_{con} = \sum_{i=1}^{N} \left(\tau_{pos} \left(\left| \mathrm{d}\left(\boldsymbol{p}_i, \boldsymbol{p}_j\right) - \mathrm{d}\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right) \right| \right) + \tau_{neg} \sum_{k=1}^{2N} \mathbb{1}_{[k \neq i,j]} \left(\left| \mathrm{d}\left(\boldsymbol{p}_i, \boldsymbol{p}_k\right) - \mathrm{d}\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right) \right| \right) \right) \tag{4.2}$$

where $\boldsymbol{z}_i$, $\boldsymbol{z}_j$ and $\boldsymbol{z}_k$ denote the predicted pose representations, and $\boldsymbol{p}_i$, $\boldsymbol{p}_j$ and $\boldsymbol{p}_k$ denote the ground-truth pose representations for the anchor, positive and negative samples in the batch. The objective of this loss function is to minimize the distance between the positive pairs and maximize the distance between the negative pairs, in

alignment with the pose similarity. Note that unlike traditional approaches where the distance is the same for all negative samples, the pairwise distance $\mathrm{d}(\boldsymbol{p}_i, \boldsymbol{p}_k)$ varies depending on the pose similarity.

4.3.4 Pixel Alignment Module

This module leverages differentiable rendering to achieve precise pixel-level alignment between the predicted mesh and the ground-truth segmentation. The alignment is enforced via a projected mask loss, which encourages the network to predict pose, shape, and camera parameters with high accuracy. This, in turn, enhances the fidelity and reliability of the overall estimation process.

4.4 Experiments

Datasets. For whole-body training, we use Human3.6M (H36M) [43], COCO-WholeBody [191] (extended from MSCOCO [45]), and MPII [48], with 3D pseudo-ground truth generated via NeuralAnnot [222]. Hand-specific training employs FreiHAND [223], InterHand [224], and COCO-WholeBody Hands [191]. For face-specific training, we adopt FFHQ [225], BUPT [226], and AffectNet [227]. Evaluation datasets are selected per task: 3DPW [65] for 3D body, FreiHAND [223] for 3D hand, and Stirling [87] for 3D face. For 3D whole-body evaluation, we use EHF [5] and AGORA [70]. Qualitative results are also presented on the MSCOCO validation set.

Metrics. We evaluate 3D joint accuracy using Mean Per Joint Position Error (MPJPE) and 3D mesh accuracy using Mean Per Vertex Position Error (MPVPE). Both are computed as the average Euclidean distance (in *mm*) between predicted and ground-truth joints or vertices after aligning the root joint translation. The pelvis is used as the root for whole-body and body, the wrists for hands, and the neck for face. Procrustes-aligned variants (PA-MPJPE and PA-MPVPE) additionally normalize for rotation and scale. For hand evaluation, we report the mean error across left and right hands.

TABLE 4.1:]

Evaluation of the Hand subnetwork on FreiHAND.

Method	PA-PVE ↓	PA-MPJPE ↓	F-Scores ↑
*** Hand-only**			
FreiHAND [223]	10.7	-	0.529/0.935
Pose2Mesh [12]	7.8	7.7	0.674/0.969
I2L-MeshNet [56]	7.6	7.4	0.681/0.973
METRO (HR64) [59]	**6.7**	**6.8**	0.717/0.981
*** Whole-body**			
ExPose [85]	11.8	12.2	0.484/0.918
Zhou et al. [218]	-	15.7	-/-
FrankMocap [86]	11.6	9.2	0.553/0.951
PIXIE [87]	12.1	12	0.468/0.919
Hand4Whole † [13]	7.7	7.7	0.664/0.971
HMR [2]	8.6	8.9	0.605/0.963
PyMAF [57]	8.1	8.4	0.638/0.969
PyMAF † [57]	7.5	7.7	0.671/0.974
`RoboSMPLX`	**7.3**	**7.5**	0.683/0.976
`RoboSMPLX` †	**7.1**	**7.4**	0.688/0.978
`RoboSMPLX` **(HR64)**	**6.7**	**6.9**	0.715/0.981

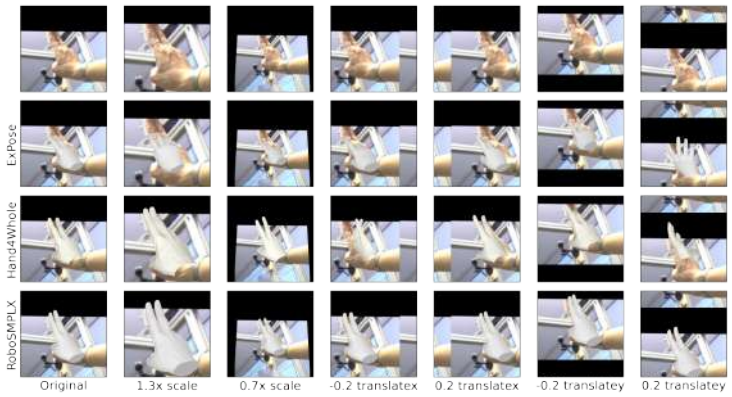
TABLE 4.2: **Evaluation of the Body subnetwork on 3DPW**.

Method	PA-MPJPE ↓	MPJPE ↓	PVE ↓
HMR (Res50) [2]	76.7	130	-
GraphCMR (Res50) [15]	70.2	-	-
SPIN (Res50) [3]	59.2	96.9	116.4
HMR-EFT (Res50) [14]	54.3	-	-
ROMP (Res50)	53.5	89.3	105.6
PARE (Res50) [4]	52.3	82.9	99.7
PARE (HR32) [4]	50.9	82	97.9
PyMAF (Res50)[57]	49.0	79.7	94.4
PyMAF (HR48) [57]	**47.1**	**78.0**	91.3
Baseline (Res50)	52.4	85.2	103.6
`RoboSMPLX` (Res50)	49.8	80.8	96.7
Baseline (HR48)	50.3	84.5	101.5
`RoboSMPLX` (HR48)	48.5	80.1	95.2

4.4.1 Benchmarking Results

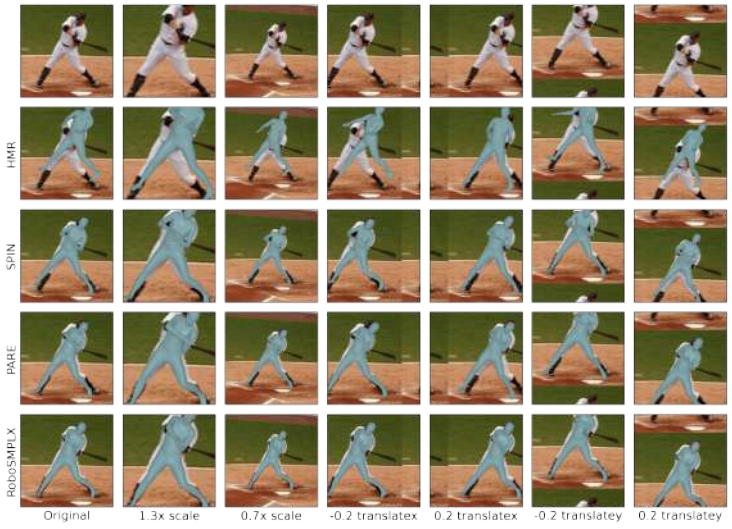
We show qualitative comparisons of `RoboSMPLX`'s Hand (Figure 4.8(a)), Face (Figure 4.8(b)) and Body (Figure 4.8(c)) subnetwork to existing models under different positional augmentations.

In general, `RoboSMPLX` s' subnetworks demonstrate better pixel alignment and are less sensitive to changes in scale and alignment.



(a) **Hand subnetwork on FreiHAND test set.** Comparison of ExPose [85], Hand4Whole [13], and `RoboSMPLX` under various augmentations.

(b) **Face subnetwork on AffectNet val set.** Comparison of ExPose [85] and `RoboSMPLX` under various augmentations.



(c) **Body subnetwork on COCO validation set.** Comparison of HMR [2], SPIN [3], PARE [4], and `RoboSMPLX` under various augmentations.

FIGURE 4.8: **Qualitative comparison of hand, face, and body subnetworks of `RoboSMPLX` under various augmentations on respective datasets.**

TABLE 4.3: **Evaluation (Face subnetwork)**.

Method	LQ Mean(mm) ↓	HQ Mean(mm) ↓
ExPose [85]	2.27	2.42
ExPose †	2.46	2.38
`RoboSMPLX`	**2.12**	**2.08**
`RoboSMPLX` †	**2.12**	**2.10**

TABLE 4.4: **PA-PVE/PVE errors of the Hand subnetwork under different positional augmentations.**

	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
Hand4Whole [13]	7.47/ 15.70	8.51/ 21.58	8.38/ 20.36	8.74/ 22.51	8.48/ 19.85	7.73/ 16.44	7.78/ 17.00
`RoboSMPLX`	**7.24/ 15.23**	**7.27/ 15.62**	**7.36/ 15.59**	**7.28/ 15.50**	**7.34/ 15.50**	**7.49/ 15.90**	**7.45/ 16.51**

TABLE 4.5: **Ablation of different modules on Body subnetwork. Results are trained on EFT-COCO and tested on 3DPW test set.**

	loss	representation	PA-	MPJPE
Baseline (HMR)	-	-	60.8	96.2
LF (all)	-	-	56.7	105.7
LF (all), L_{con}	L1	pose	55.9	90.9
LF (all), L_{con}	MSE	pose	58.5	93.9
LF (all), L_{con}	SmoothL1	pose	56.6	92.5
LF (all), L_{con}	L1	pose(rot6d)	58.9	95.0
LF (all), L_{con}	L1	pose + go	76.8	118.9
LF (all), L_{con}, +ve	L1	keypoints	**55.4**	**90.56**

Hand Subnetwork. Table 4.1 reports the performance of the Hand subnetwork against both hand-only and whole-body baselines. When trained solely on FreiHAND (e.g., PIXIE, Hand4Whole, PyMAF) or with mixed datasets (Hand4Whole †, PyMAF †)[0] using the same backbone, `RoboSMPLX` consistently surpasses its whole-body counterparts. Previous studies [42, 56] have shown that whole-body approaches often rely on parametric hand mesh models, which tend to underperform compared to the non-parametric mesh outputs in recent hand-only methods [56, 59]. Despite this well-documented gap, `RoboSMPLX` not only narrows but exceeds it, outperforming mesh-based methods and achieving results on par with the state-of-the-art METRO, when matched for backbone capacity (HRNet-64). Robustness under positional perturbations is assessed in Table 4.4, where `RoboSMPLX` demonstrates substantially lower errors than Hand4Whole under translation and scaling variations. Additional qualitative examples are provided in [215].

[0]† indicates training with additional datasets in the subsequent evaluations and tables.

TABLE 4.6: **3DRMSE errors of the Face subnetwork under different positional augmentations.**

	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
ExPose [85]	2.27	2.38	2.29	2.46	2.30	2.46	2.27
`RoboSMPLX`	**2.12**	**2.20**	**2.17**	**2.13**	**2.18**	**2.24**	**2.10**

Body Subnetwork. Table 4.2 compares Body subnetwork results on the 3DPW test set. `RoboSMPLX` achieves competitive performance relative to other SMPL-based methods, though it is not the top-performing approach. It is important to note that reported results in the literature are often obtained under differing training settings, including variations in backbone initialization, training datasets, and optimization strategies [181], which can significantly affect performance. In contrast, our method is trained under a controlled and consistent setup to enable fair ablation analysis.

The ablation results in Table 4.5 show that each proposed module contributes meaningful improvements over this baseline. Furthermore, RoboSMPL-X is primarily designed to improve robustness in whole-body estimation, particularly for hands and face under challenging conditions, rather than optimizing body-only benchmarks. This design trade-off explains why the body subnetwork does not always achieve the highest performance, while still leading to improved overall whole-body results.

Face Subnetwork. Table 4.3 evaluates the Face subnetwork on the Stirling3D test set. With identical training data, `RoboSMPLX` outperforms ExPose. Notably, ExPose experiences degraded accuracy when trained on multiple datasets, whereas `RoboSMPLX` maintains consistently low errors. Qualitative examples in Fig. 4.8(b) illustrate strong generalization to in-the-wild conditions. Robustness tests in Table 4.6 confirm that `RoboSMPLX` is less sensitive to translation and scale perturbations. Additional visualizations are available in [215].

Whole-body Network.

We further assess the whole-body network on two benchmarks: the EHF validation set and the AGORA test set (Table 4.7). On EHF, `RoboSMPLX` outperforms existing full-body methods, with notable gains in hand and face estimation, and demonstrates robustness under positional perturbations (Table 4.8).

On AGORA, performance is comparatively lower. This can be attributed to two factors. First, AGORA contains more severe multi-person interactions and

TABLE 4.7: **Evaluation of wholebody network on EHF and AGORA test set.**

Method	EHF						AGORA					
	PVE ↓			PA-PVE ↓			PVE ↓				N-PVE ↓	
	WB	H	F	WB	H	F	WB	B	F	LH/RH	WB	B
ExPose [85]	77.1	51.6	35	54.5	12.8	5.8	217.3	151.5	51.1	74.9/71.3	265	184.8
PIXIE [87]	89.2	42.8	32.7	55	11.1	**4.6**	191.8	142.2	50.2	49.5/49.0	233.9	173.4
Hand4Whole [13]	76.8	39.8	26.1	50.3	10.8	5.8	135.5	90.2	41.6	46.3/48.1	144.1	96.0
OSX (ViT-L) [84]	**70.8**	53.7	26.4	**48.7**	15.9	6	**122.8**	**80.2**	**36.2**	45.4/46.1	**130.6**	**85.3**
Ours	73.7	**34.9**	**17.8**	49.7	**10.0**	**4.6**	132.3	85.0	39.4	**45.3/46.1**	138.2	91.5

occlusions, which increase the likelihood of incorrect subject association during inference (Fig. 4.9). Second, reported results across methods are not always directly comparable, as they may differ in training data composition, pretraining strategies, and detection pipelines. In our case, `RoboSMPLX` is trained under a unified and controlled setup without dataset-specific tuning, prioritizing robustness across conditions rather than optimizing for a single benchmark.

Qualitative comparisons (Fig. 4.10) show that once the target subject is correctly localized, `RoboSMPLX` achieves superior pixel-level alignment and more precise hand and face reconstructions, indicating that its whole-body accuracy remains competitive in challenging scenarios once the detection bottleneck is removed.

FIGURE 4.9: **Visualisation of samples with high errors at train time.** Red and green vertices indicate the target person model's predictions respectively.

4.4.2 Ablation Studies

Location features. Table 4.12 presents the ablation of modules for the Hand subnetwork (the Body subnetwork ablation is in Table 4.5). The baseline is trained

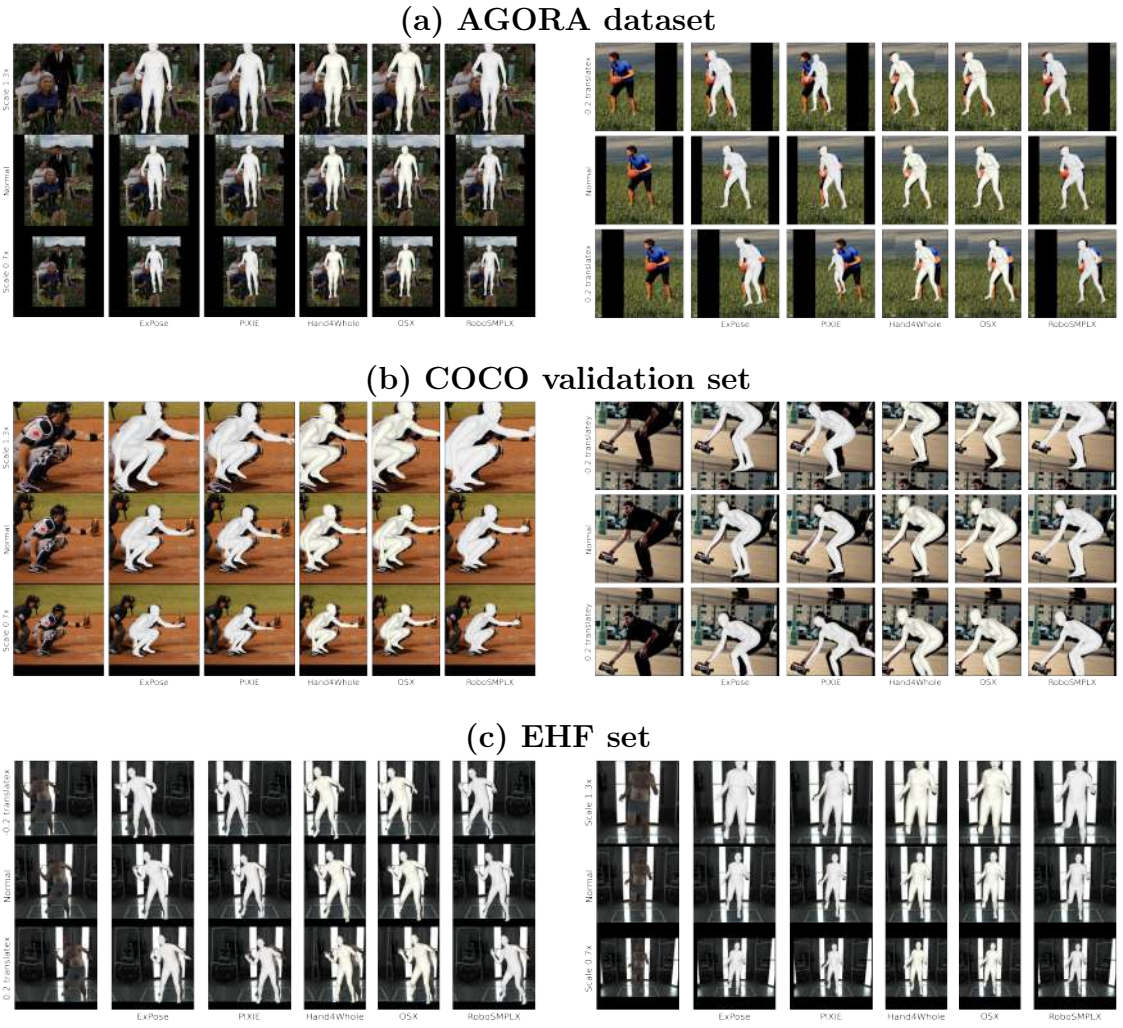


FIGURE 4.10: **Visualization of Expose [85], PIXIE [87], Hand4Whole [13], OS-X [84], and `RoboSMPLX` under different scales and alignments on (a) AGORA, (b) COCO, and (c) EHF datasets. Each row corresponds to one dataset.**

with random scale factor 0.2 and bounding-box jitter 0.2; applying stronger augmentations (*strongaug*) reduces accuracy, likely due to domain shift. Hand4Whole [13] extracts joint features via positional pose-guided pooling (PPP) for pose estimation, while relying only on backbone features for shape and camera prediction. Our approach explicitly learns both sparse and dense location cues, including part silhouettes, and uses these location features ("LF") for pose, shape, and camera estimation. Incorporating LF for all three predictions ("LF (all)") yields further performance gains, lowering both joint and vertex errors of the regressed mesh.

TABLE 4.8: **Wholebody, Hand and Face PA-PVE errors under different positional augmentations.**

	Method	Normal	Transx +0.2x	Transx -0.2x	Transy +0.2y	Transy -0.2y	Scale 1.3x	Scale 0.7x
Hands	ExPose [85]	14.39	17.36	17.86	14.93	17.21	14.15	14.56
	PIXIE [87]	14.68	15.05	16.11	15.32	15.85	14.52	14.79
	Hand4Whole [13]	10.83	11.15	11.34	10.50	13.70	10.77	11.25
	OSX [84]	15.97	16.42	16.55	16.94	17.86	15.91	17.24
	RoboSMPLX	**10.00**	**10.37**	**10.21**	**10.16**	**12.49**	**9.98**	**10.19**
Face	ExPose [85]	6.34	10.28	6.71	8.17	6.43	6.24	6.24
	PIXIE [87]	5.63	6.67	6.94	6.53	6.94	5.84	5.84
	Hand4Whole [13]	5.81	5.88	5.91	5.74	5.93	5.76	5.76
	OSX [84]	6.09	6.03	6.09	5.83	5.96	5.92	5.92
	RoboSMPLX	**4.65**	**5.10**	**5.38**	**4.75**	**5.30**	**4.77**	**5.22**
Wholebody	ExPose [85]	54.82	61.64	65.98	65.03	65.98	54.03	59.23
	PIXIE [87]	54.85	66.16	69.26	64.83	69.26	56.28	60.31
	Hand4Whole [13]	50.37	59.10	67.85	64.64	67.85	48.10	55.28
	OSX [84]	**48.79**	**51.09**	55.96	95.97	**55.96**	**47.35**	**50.89**
	RoboSMPLX	49.79	52.46	**53.62**	**61.65**	63.99	47.90	51.39

TABLE 4.9: **Ablation of contrastive learning methods and loss.**

Scale factor	Mean ↓	Std ↓
SimCLR	0.227	0.0915
SimCLR (+ pose-variant aug.)	0.230	0.0911
SimCLR (+ background aug.)	0.222	0.0959
SimCLR (+ L_{con})	0.164	0.0772
HMR	0.140	0.0823
HMR (+ L_{con})	0.124	**0.0624**
HMR (+ L_{con}, +ve samples)	**0.119**	0.0679

TABLE 4.10: **Ablation of different representation for contrastive loss.**

Representation	PA-↓	MPJPE↓	PA ↓	PVE↓
baseline	7.49	15.51	7.46	15.59
pose	8.11	15.81	7.67	16.08
go + pose	7.71	14.98	7.54	14.91
keypoint	7.48	15.01	7.32	15.29
pose, +ve	7.45	14.94	7.20	**14.77**
keypoint, +ve	**7.31**	**14.62**	**7.18**	15.01

Takeaway. Explicitly learning both sparse and dense location cues, and reusing them as features for pose, shape, and camera prediction, is more effective than sampling features at predicted joint coordinates. Location-aware features propagate consistently along the kinematic chain and reduce both joint and vertex errors.

Contrastive loss. We first validate that prior contrastive SSL methods [88–90] are suboptimal for learning pose- and shape-aware embeddings. As illustrated in

TABLE 4.11: **Ablation of augmentation +ve samples, using pose rotation as representation.**

Augmentation	PA-↓	MPJPE↓	PA- ↓	PVE↓
baseline (no +ve)	8.11	15.81	7.67	16.08
color	7.42	15.01	7.18	14.94
pose	8.59	16.96	8.15	17.21
location	7.80	15.98	7.46	15.56
color + location	**7.45**	**14.94**	**7.20**	**14.77**

TABLE 4.12: **Ablation of different modules on Hand subnetwork. Results are trained and evaluated on FreiHAND.**

	Supervision	PA-↓	MPJPE↓	PA-↓	PVE↓
Base (R50)		8.06	16.78	7.85	16.71
Base (R50) + Strongaug		8.47	17.01	8.11	16.17
Base (DR54)		7.8	15.57	7.67	15.72
Base (DR54)	L_{KS}	7.68	15.8	7.62	16.29
PPP [13]	L_{KS}	7.65	15.93	7.56	16.37
LF	L_{KS}	7.52	15.84	7.56	16.15
joints	L_{KS}	7.86	15.92	7.75	16.24
LF (all)	L_{KS}	7.49	15.51	7.46	15.59
LF (all) + L_{con}	L_{KS}	7.48	15.01	7.32	15.29
LF (all) + L_{con}, +ve	L_{KS}	7.42	14.88	7.16	14.57
LF (all)	L_{KS}, L_{segm}	7.44	14.92	7.58	15.30
LF (all)	L_{KS}, L_{segm}, L_{proj}	7.36	14.38	7.53	15.05
LF (all)+ L_{con}, +ve	L_{KS}, L_{segm}, L_{proj}	**7.33**	**14.59**	**7.02**	**14.11**

Fig.4.11, retrieval based on SSL-learned embeddings (top-5 matches) tends to emphasize background similarity rather than pose similarity. Quantitatively, Table4.9 reports the estimation errors of the top-1 retrieved pose (COCO-train) relative to the query pose (COCO-test). SimCLR yields higher mean errors than supervised HMR, consistent with the findings of [90] that SSL representations transfer poorly to pose and shape estimation. In contrast, `RoboSMPLX` 's incorporation of a contrastive loss with positive samples ("HMR + L_{con}, +ve") encourages embeddings of the same subject under varied augmentations to converge, improving robustness.

Table 4.10 evaluates three pose representations (Sec. 4.3.3): (1) *pose* – concatenated global orientation and rotational pose; (2) *go+pose* – orientation and pose as separate entities; (3) *keypoint* – root-aligned 3D joints regressed from the body model. The joint-based representation consistently outperforms pose-based ones, as it encodes both shape and pose in a normalized space. Representations using relative rotations

FIGURE 4.11: **Left: Query image from the EFT-COCO-Test set, Right: Retrieved image from the EFT-COCO-Train set ordered in descending embedding similarity.**

are less effective, likely because small rotation errors can accumulate into substantial joint discrepancies. Incorporating positive pairs ("pose, +ve" and "keypoint, +ve") further improves performance, confirming that similarity-preserving augmentations enhance the learned feature space.

Table 4.11 compares augmentation strategies. While prior works [88, 89] included pose-variant augmentations (rotation, flipping), we observe these are detrimental, as they alter global orientation. In contrast, location-variant and color-variant

FIGURE 4.12: **Comparison of keypoints and pose representations.**

augmentations—individually or combined—consistently improve performance over the baseline.

Figure 4.12 compares retrieval based on pose similarity (rot6d) and keypoint similarity. High keypoint similarity usually corresponds to high pose similarity, but the reverse does not always hold: small differences in joint rotations can accumulate into significant mismatches in keypoint positions. This explains why joint-based representations yield better results in Table 4.10, as they retrieve samples that are more closely aligned with the query image both visually and geometrically.

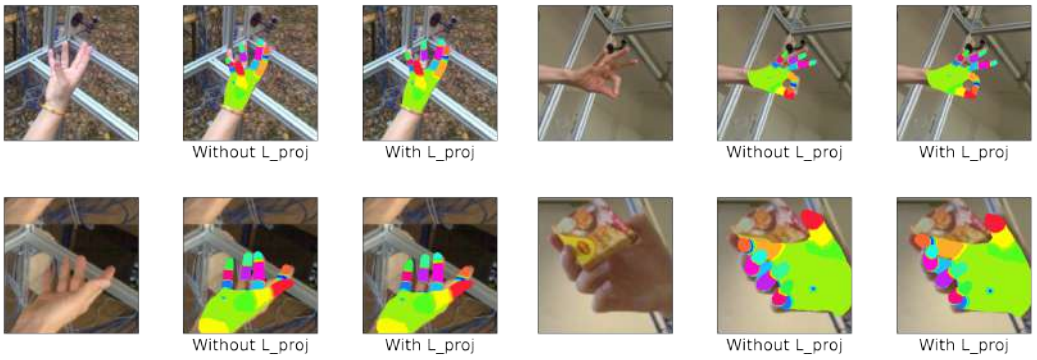


FIGURE 4.13: **(C) Visualisation from training with and without L_{proj}.**

Takeaway. A pose- and shape-aware supervised contrastive objective, with positive samples generated by location- and color-variant augmentations, yields embeddings that capture pose similarity rather than background similarity. This encourages augmentation-invariant features, which keep the pose and shape estimates accurate under imperfect part crops.

Pixel alignment. Table 4.12 shows that incorporating differentiable rendering and applying the projected segmentation loss (L_{proj}) in `RoboSMPLX` reduces both PVE and MPJPE errors. These improvements stem from more accurate learning of body and camera parameters, leading to better alignment between the rendered 3D mesh and the input image. Conventional metrics such as PVE and MPJPE are computed after root alignment and therefore do not fully capture how well the predicted mesh aligns with the image once reprojected to 2D. This limitation is further compounded by the lack of ground-truth camera parameters in most pose and shape estimation datasets, meaning that camera supervision is typically sparse and indirect, provided only through aligning projected joints with ground-truth 2D keypoints. The pixel alignment strategy addresses this by enforcing denser supervision via L_{proj}, which improves both the estimation of camera parameters and the overall quality of the re-projection. Qualitative results (Fig. 4.13) show that training with L_{proj} leads to visibly better alignment of the projected mesh vertices with the image compared to training without it.

Takeaway. Differentiable rendering with a projected segmentation loss (L_{proj}) provides dense 2D supervision. It simultaneously improves camera estimation and the visual alignment of the rendered mesh to the image, a quality under-reported by standard numerical metrics.

4.5 Conclusion

In this chapter, we introduce a new framework `RoboSMPLX` to advance the field of whole-body pose and shape estimation. Our approach enhances the whole-body pipeline by improving the accuracy of part crop localization and equipping part subnetworks with the robustness needed to handle imperfect crops while still producing reliable results. This is achieved through three key innovations: (1) precise subject localization via explicit learning of both sparse and dense part predictions, (2) robust feature representation using supervised contrastive learning, and (3) accurate pixel-level alignment of the outputs through differentiable rendering.

Future Work. Several promising research directions remain. First, our current training strategy does not actively select hard negative samples. Investigating hard negative mining—such as deliberately including visually similar poses within the same batch—may improve the discriminative power of the learned features. Second, although we avoid augmentations that alter global orientation (e.g., flipping or rotation) due to their negative impact, we have not fully studied the individual and combined effects of other augmentations. Future efforts could explore methods for automatically selecting and composing augmentations to maximize performance.

Within the broader thesis narrative, this chapter provides a stronger body-level foundation for the subsequent move to clothed and disentangled avatar modeling in Chapter 5. While the proposed framework improves robustness under imperfect crops, challenges such as loose clothing and heavy occlusion remain, and seed the future directions outlined in Chapter 7.

Chapter 5

Disentangled 4D Human Generation and Animation from a Single Image

5.1 Introduction

Building on the improved robustness in whole-body pose and shape estimation from Chapter 4, this chapter turns to the next stage of the thesis pipeline: moving from body recovery to clothed and disentangled avatar modeling. We introduce Disco4D, which represents clothing, hair, and accessories as separable Gaussian layers over a parametric SMPL-X body, enabling more flexible reconstruction, editing, and animation. A remaining gap motivates the work in Chapter 6: canonical multi-view supervision for consistent asset recovery.

The development of high-fidelity 3D digital humans is increasingly important across a variety of augmented and virtual reality applications. To streamline the creation of these digital avatars from easily accessible in-the-wild images, a multitude of research efforts have been made on reconstructing 3D clothed human models from a single image [18–24, 24–27]. These works predominantly focus on the simultaneous reconstruction of the human body and clothing. Unfortunately, these works have inherent limitations, and integrating them into applications that require virtual try-on or avatar customization poses significant challenges. This is primarily because the models are rendered as single-layer, non-animatable meshes where distinct attributes (e.g., hair, clothing, accessories) are merged into one

The work in this chapter has been published in [228].

continuous surface, with underlying layers completely obscured and self-contact areas inseparably connected. Such limitation complicates the re-animation and dynamic customization tasks. Existing works that perform layered reconstruction [144, 145] rely on self-rotating video inputs with extensive frames and viewpoints and involve substantial processing times.

To address these issues, we propose Disco4D, a novel 4D clothed human reconstruction method that *distinctly separates the human body from clothing elements* from a single image. It supports human animation as the 4th dimension, which cannot be realized by prior static 3D reconstruction works [18–25]. To achieve this, it employs the SMPL-X [5] parametric model to represent the human body, capitalizing on its efficacy in capturing body structure and kinematics. Conversely, clothing, along with dynamic and variable elements such as hair and accessories, is represented using Gaussian models, which are able to model the large variability in clothing. By binding Gaussians to a SMPL-X model and fixing it during the training phase, Disco4D ensures the integrity of the body while focusing the learning process on the appearance aspects. To model occluded portions not visible in the input image, diffusion models are used to enhance the 3D generation process. Moreover, Disco4D includes an identity grouping mechanism for the Gaussians, which is instrumental in maintaining the separability and individuality of each clothing asset.

The independent reconstruction of clothing and body offers several advantages. (1) *Enhanced reconstruction fidelity.* The SMPL-X body serves as a stable anchor for the clothing to conform to. By isolating the focus to learn clothing Gaussians, we achieve a more refined geometry and intricate detailing in the clothed model. (2) *Fine-grained categorization and extraction of clothing items.* Disco4D is able to separate clothing Gaussians into their respective categories, which is crucial for the recovery and utilization of individual clothing assets. (3) *Extensive editing capabilities.* Disco4D supports different editing functions, including the removal of specific items, inpainting (altering color or material), and other modifications. Such rich editing options allow for precise adjustments to individual assets without inadvertently affecting adjacent elements. This level of control is particularly beneficial in applications requiring detailed customization, such as virtual fashion design and digital content creation. (4) *Improved animation capabilities.* The body Gaussians adhere to the deformations dictated by the SMPL-X model, while clothing Gaussians conform to the underlying body movements but also exhibit

behaviors true to their material characteristics. The disentangled deformation allows for nuanced adjustments to clothing behavior in response to complex body movements, thereby elevating the quality of clothed human animation.

5.2 Methodology

5.2.1 Preliminaries

3D Gaussian Splatting employs explicit 3D Gaussian points as its primary rendering entities. A Gaussian at position x is modeled as a function $G(x) = e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$, where μ denotes the spatial mean and Σ is the covariance matrix. Each Gaussian additionally has its own rotation r, scaling s, opacity α, and a view-dependent color c parameterized via spherical harmonic coefficients f. Rendering is performed by projecting these 3D Gaussians into the 2D image plane through a splatting operation. The projection uses the camera projection matrix, and the corresponding 2D covariance is approximated as: $\Sigma' = J_g W_g \Sigma W_g^T J_g^T$, where W_g is the view transformation, and J_g is the Jacobian of the affine approximation for perspective projection. The final pixel color is obtained through alpha-blending of N layered 2D Gaussians from front to back $C = \sum_{i \in N} T_i \alpha_i c_i$, with $T_i = \prod_{j=1}^{i}(1-\alpha_j)$.

Here, the opacity α is obtained by multiplying a learnable density term γ with the contribution of the projected covariance Σ' relative to the pixel position in image space. For optimization, the 3D covariance Σ is parameterized using a quaternion q for rotation and a scale vector v.

SMPL-X parameterization [5] extends the original SMPL body model [1] by incorporating detailed face and hand deformations to better model more expressive human motions. **SMPL-X** expands **SMPL** joint set by including additional joints for facial features, toes and fingers, allowing a more precise modeling of complex body motions. **SMPL-X** is represented by a function $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta| \times |\theta| \times |\psi|} \rightarrow \mathbb{R}^{3N}$, where $\theta \in \mathbb{R}^{3K}$ represents the pose (with K being the number of body joints), $\beta \in \mathbb{R}^{|\beta|}$ represents body shape, and $\psi \in \mathbb{R}^{|\psi|}$ captures facial expressions. Further details can be found in [5].

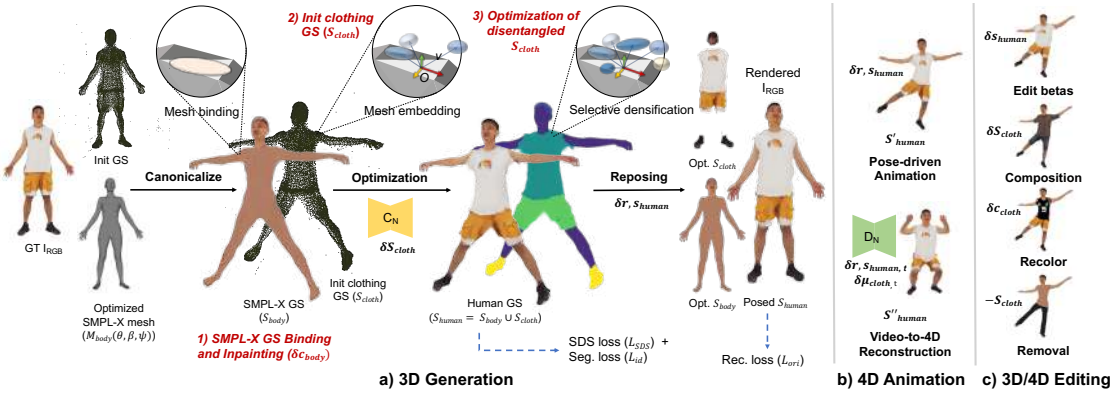


FIGURE 5.1: **Framework Overview of Disco4D. (a) 3D Generation** utilizes a single image to obtain disentangled body and clothing Gaussians. Body, face and hand poses are refined to be pixel-aligned. For faster initialization, clothing Gaussians and visual hull are obtained with Gaussian Reconstruction Models. These clothing Gaussians are embedded to SMPL-X mesh and adopt the local coordinate system of the triangle. Subsequently, the iterative optimization process (pruning, identity encoding and densifying) separates the body and garments. The learned identity encodings guide the densification of the clothing Gaussians. **(b) 4D Animation** is achieved by either direct driving of SMPL-X poses or leveraging video to learn extra clothing deformation (refer to Figure 5.2 for more details). Various **(c) 3D/4D Editing** operations can be performed with our disentangled representation.

5.2.2 Overview

Given a single image, Disco4D generates animatable 3D clothed human avatars in a bottom-up manner, facilitating natural separability. Our generated 3D clothed avatars, denoted as S_{human}, are represented as the concatenation of S_{body} and S_{cloth}. Inspired by prior works [115, 116], S capitalizes on Gaussian representations:

$$S = G(\mu, r, s, \alpha, c, e), \tag{5.1}$$

where μ, r, s, α, c and e denote *positions*, *rotation*, *scaling*, *opacity*, *spherical harmonics coefficients* and *identity encoding*, respectively. Different from traditional Gaussian representations, we add identity encoding e to associate each Gaussian with its clothing category.

Figure 6.1 depicts our framework. We start by generating colored SMPL-X Gaussians representing the body beneath clothing (Sec. 5.2.3). We obtain a visual hull for canonicalization and refine Gaussian predictions to align and envelop the SMPL-X mesh (Sec. 5.2.4). Next, we iteratively optimize canonical clothing Gaussians

external to the SMPL-X mesh (Sec. 5.2.5). Lastly, we showcase the animation and editing of generated clothed avatars (Sec. 5.2.6). Notably, we leverage diffusion models to refine textures during 3D generation (Sec. 5.2.5) and extrapolate unseen views during 4D animation (Sec. 5.2.6).

5.2.3 SMPL-X Gaussians

Given an image, we first estimate coarse SMPL-X parameters with an off-the-shelf model [229], and then refine coarse predictions by fitting on 2D keypoints and clothing segmentation masks, obtaining pixel-aligned SMPL-X parameters (β, θ, ψ).

Mesh Binding. To convert the SMPL-X [5] mesh $M(\beta, \theta, \psi)$ into Gaussians S_{body} for rendering, flat 3D Gaussians are bound to each mesh triangle, similar to SuGaR [230]. Gaussian means μ_{body} are computed using predefined barycentric coordinates, while Gaussian rotations r_{body} derive from surface normals. The initial scaling s_{body} ensures dense mesh coverage, with the last axis set to 0.1 for a uniformly thin surface. For color representation beneath clothing, opacity α_{body} is set to 1.0, with spherical harmonics c_{body} optimized for each Gaussian. Visible skin color is supervised, while occluded skin color aligns with visible regions. A fixed label e_{body} is assigned for rendering, remaining unchanged during training. When optimizing clothing Gaussians S_{cloth}, SMPL-X Gaussians S_{body} parameters stay fixed, preserving the body structure while allowing flexible learning for clothing.

5.2.4 Initialization of Clothing Gaussians

Cloth styles are diverse, making proper initialization crucial for effective clothing modeling. In synchronization with estimating SMPL-X, we first employ the Video Diffusion Model [118] to estimate multi-view images. Subsequently, we leverage Gaussian Reconstruction Models [116] to obtain initial 3D Gaussians and their corresponding visual hull. Yet, the reconstructed 3D outputs often suffer from geometric inaccuracies, such as incorrect poses due to pose ambiguity or missing limbs. To address this, we refine the coarse visual hull to ensure it accurately aligns with and overlays the SMPL-X mesh and encapsulates a good geometry for the clothed figure. With SMPL-X aligned visual hull, we derive the refined Gaussians

by adopting properties from their nearest neighbors. The refined visual hull and Gaussians are then canonicalized for the optimization phase.

Mesh embedding. Each 3D clothing Gaussian is embedded on a triangle of the canonical mesh, defining its position in both canonical and posed spaces. The mean vertex position $\mathbf{O}$ serves as the origin of the local coordinate system, with the Gaussian positioned by an offset vector $\mathbf{v} = \sigma\mathbf{i} + \beta\mathbf{j} + \gamma\mathbf{k}$, where σ, β, and γ are the components of the displacement vector along the tangent $\mathbf{i}$, bitangent $\mathbf{j}$, and normal $\mathbf{k}$. Unlike SplattingAvatar [231], which displaces Gaussians along the normal, our approach allows embedding to the most suitable triangle rather than the nearest one. For example, hair Gaussians are tagged to head faces instead of the nearest face for reposing [231, 232] (Figure 5.12). In animation, the Gaussian rotates with its embedded triangle face (δr), while scaling (δs) is adjusted dynamically based on changes in edge lengths. During optimization, Gaussian and embedding parameters ($\mathbf{O}$, $\mathbf{v}$, δr, and δs) are jointly updated.

5.2.5 Optimization of Separable Gaussians

With the SMPL-X Gaussian and initialized clothing Gaussian, we aim to optimize canonical clothing Gaussians S_{cloth} outside the SMPL-X mesh. This involves three steps: **1)** we use Signed Distance Function (SDF) loss and pruning to discourage and remove Gaussians that reside within the body; **2)** we introduce *identity encoding* e to attach a clothing label for each clothing Gaussian, by lifting multi-view 2D segmentations of the target object onto the 3D Gaussians; and **3)** guided by e_{body} and e_{cloth}, we selectively densify only the relevant clothing points while ignoring body points. Once the disentangled clothing is obtained, we use SDS loss to in-paint high-resolution texture from the reference image to individual clothing Gaussians, thereby enriching the details of unseen regions.

SDF Loss and Pruning. In reality, the clothing is always external to the body. During refinement, we ensure that the clothing Gaussians are positioned externally to the SMPL-X mesh by applying the SDF loss and a pruning strategy. Specifically, the SDF loss $\mathcal{L}_{sdf}$ penalizes any new densified Gaussians that intrude into the space of the SMPL-X mesh, ensuring that the clothing Gaussians consistently remain outside the body's surface. Pruning is applied at fixed intervals to reinforce this

separation, and systematically remove any Gaussians located within the SDF of the SMPL-X mesh.

Identity encoding. To associate each Gaussian to its clothing category, we introduce *Identity Encoding (e)*, a learnable and compact vector of length 15, representing clothing categories from SegFormer [233] segmentation masks[0]. During training, the encodings are rendered into 2D segmentation masks in a differentiable manner following [234]. For classification, we apply a softmax to the rendered features E_{id} and use cross-entropy loss $\mathcal{L}_{2d}$ for $(K+1)$-category classification. An unsupervised 3D regularization loss $\mathcal{L}_{3d}$ promotes spatial consistency among the top k-nearest 3D Gaussians' Identity Encodings. Consequently, the overall identity loss is $\mathcal{L}_{id} = \mathcal{L}_{2d} + \mathcal{L}_{3d}$. Refer to Appendix in the [228] for more details.

Densification of clothing Gaussians. To learn clothing more efficiently, we perform sampling for categorical Gaussians that belong to the same clothing category and embedding. We find the k-nearest Gaussian points for the resampled points and inherit their Gaussian properties (scaling, rotation, opacity, SH properties). By selectively densifying clothing Gaussians, we only add necessary Gaussians while ignoring body Gaussians.

Anisotropy. To prevent overly-skinny kernels that point outward from the object surface under large deformations, we enforce the anisotropy of Gaussian kernels following [235]. During optimization, we employ $\mathcal{L}_{\text{ani}} = \frac{1}{|P|}\sum_{p\in P}\max\left(\frac{\max(s_p)}{\min(s_p)}, \tau\right) - \tau$, where s_p is the scalings of 3D Gaussians. This loss constrains the ratio between the major and minor axis lengths below threshold τ.

Total loss. To inpaint occluded textures, we use the $\mathcal{L}_{SDS}$ loss on the Gaussians in the canonical pose after optimizing the front view for 500 steps. Combined with the conventional 3D Gaussian Loss $\mathcal{L}_{ori}$ on image rendering, the total loss L for end-to-end optimization of clothing Gaussians via network C_N is:

$$\mathcal{L} = \lambda_{ori}\mathcal{L}_{ori} + \lambda_{id}\mathcal{L}_{id} + \lambda_{ani}\mathcal{L}_{ani} + \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{SDS}\mathcal{L}_{SDS} \tag{5.2}$$

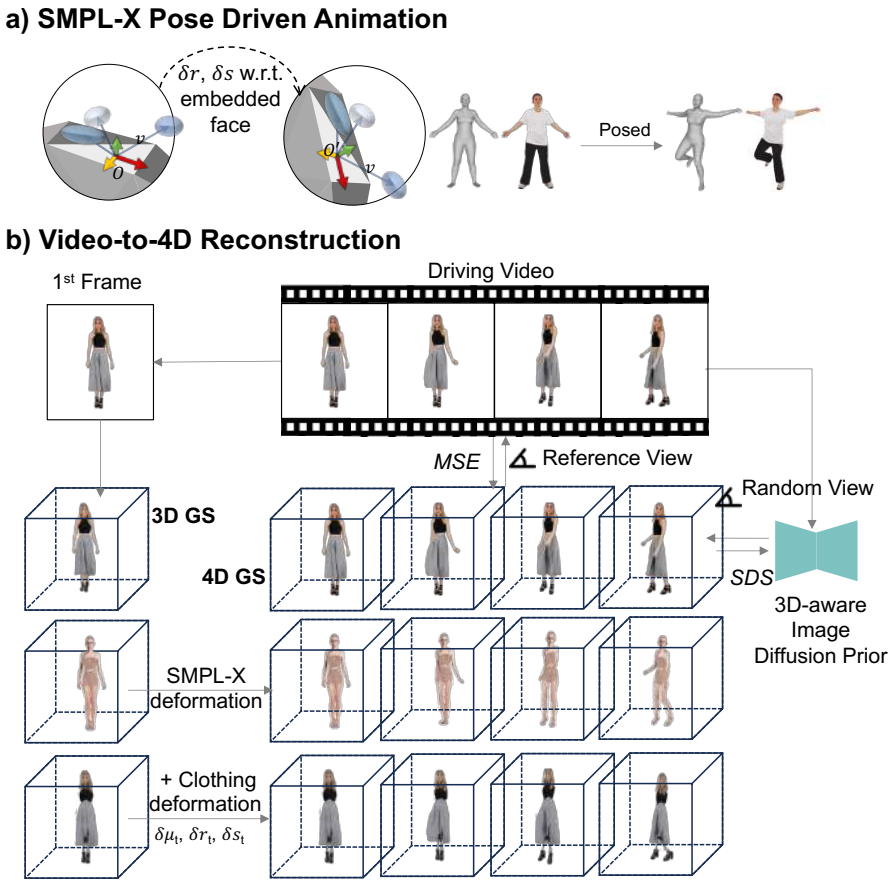


FIGURE 5.2: **4D animation** is achieved by (a) driving SMPL-X poses or (b) using video to learn additional clothing deformations. From the first frame, a static 3D disentangled GS model is generated. Pose transformations deform body and clothing Gaussians, and a deformation network is optimized to capture additional clothing deformations over time.

5.2.6 4D Human Animation and Editing

Disco4D's disentangled representation naturally supports animation and editing. The canonical Gaussians S_{body} and S_{cloth} enable separate deformations for clothing and body, ensuring realistic animation. Besides, individual clothing categories can be easily edited using image or text prompts. The learned clothing can be transferred to different body shapes and poses, for versatile customization.

Animating Gaussians. As shown in Figure 6.1, Disco4D enables animation of the canonical human Gaussian via two methods. Firstly, Gaussians can be directly driven using 3D SMPL-X sequences obtained from a motion database or

[0]Categories: 0: "Background", 1: "Hat", 2: "Hair", 3: "Sunglasses", 4: "Upper-clothes", 5: "Skirt", 6: "Pants", 7: "Dress", 8: "Belt", 9: "Left-shoe", 10: "Right-shoe", 11: "Face", 12: "Skin", 13: "Bag", 14: "Scarf"

estimated from 2D videos. Secondly, Disco4D enhances the model by learning detailed clothing dynamics from monocular videos. This disentanglement enables the focused modeling of clothing dynamics without altering the underlying human representation.

To extend static 3D Gaussians into dynamic 4D Gaussians, a deformation network is trained to predict changes in position, rotation, and scale of the reposed clothing Gaussians based on a timestamp, as described in DreamGaussian4D [173]. Unlike DreamGaussian4D [173], which learns deformations for all Gaussians, Disco4D models body Gaussians using the SMPL-X mesh, while clothing Gaussians employ posed transformations and learned deformations. The transformation is defined as $S'' = D_N(S', t)$ where D_N is the deformation network, S' is the spatial descriptions of the reposed 3D clothing Gaussian, t is the timestamp, and S'' is the spatial descriptions of the deformed and reposed 3D clothing Gaussians. Following DreamGaussian4D [173], the deformation model is initialized to predict zero deformation at the start of training to avoid divergence between dynamic and static models. The weights and biases of the final prediction heads are initialized to zero, and skip connections are introduced to enable gradient backpropagation.

To optimize the deformation field using the reference view video, we minimize the reconstruction loss $\mathcal{L}_{Ref}$ between the rendered image and video frame at each timestep. To propagate the motion from the reference view to the entire 3D model, we leverage Zero-1-to-3-XL [106] to predict the deformation of the unseen part to calculate $\mathcal{L}_{SDS}$. Despite per-frame predictions of image diffusion models, the fixed color and opacity of static 3D Gaussians help preserve temporal consistency.

Editing Clothing Gaussians. We extract the Gaussians corresponding to the specific category and edit them. This allows fine-grained editing and ensures that other Gaussians are not affected. Instead of fine-tuning all 3D Gaussians, we freeze the properties for most of the well-trained Gaussians and only adjust a small part of 3D Gaussians relevant to the target categories. For 3D object removal, we simply delete the 3D Gaussians of the editing target. For 3D object colorization by in-painting or text guidance, we reinitialise the color and tune the color (SH) parameters of the corresponding Gaussian group, while fixing the 3D positions and other properties to preserve the learned 3D geometry.

TABLE 5.1: **CLIP-embedding loss for generated humans and segmented assets, and performance (PSNR, SSIM, LPIPS) comparisons for novel poses and views on the Synbody and CloSe datasets across DreamGaussian, LGM, SHERF, and Disco4D.**

Method	SynBody							CloSe									
	CLIP				Novel View			CLIP				Novel View			Novel Pose		
	All ↑	Pants ↑	Shirt ↑	Shoes ↑	PSNR ↑	SSIM ↑	LPIPS ↓	All ↑	Pants ↑	Shirt ↑	Shoes ↑	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
DreamGaussian	0.751	0.715	0.710	0.749	13.118	0.883	0.229	0.734	0.693	0.674	0.767	20.08	0.939	0.089	-	-	-
LGM	0.807	0.724	0.747	0.760	12.884	**0.876**	0.228	0.829	0.727	0.712	0.778	**20.50**	**0.939**	**0.077**	-	-	-
SHERF	0.766	0.649	0.636	0.714	15.189	0.852	0.189	0.777	0.785	0.729	0.801	18.96	0.912	0.083	15.54	0.844	0.165
Disco4D	**0.851**	**0.784**	**0.753**	**0.801**	**15.691**	0.848	**0.185**	**0.856**	**0.858**	**0.810**	**0.842**	20.10	0.918	0.081	**17.96**	**0.851**	**0.136**

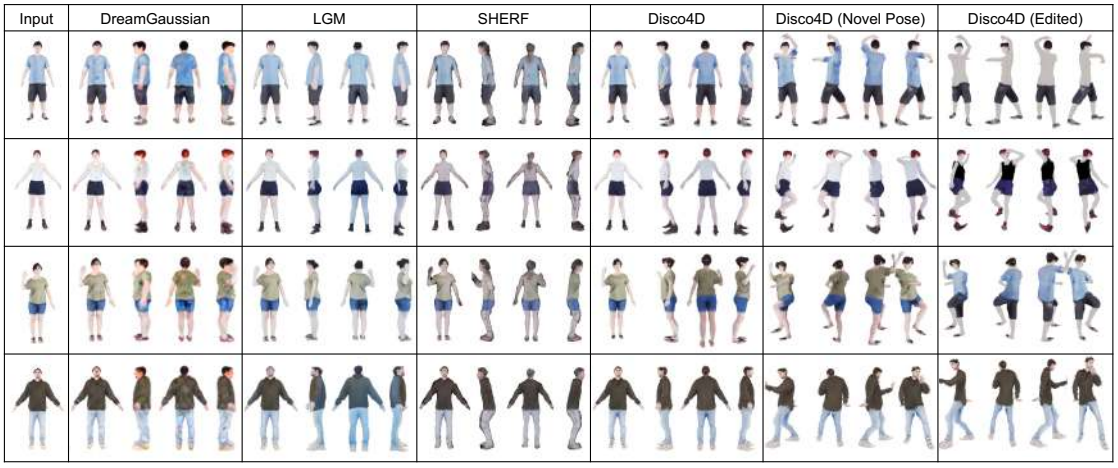


FIGURE 5.3: **Qualitative comparison of image generation across DreamGaussian, LGM, SHERF, and Disco4D.**

5.3 Experiments

5.3.1 Implementation details

The 3D generation experiments were conducted using a single 24GB RTX3090 GPU, while the 4D generation experiments utilized a single 48GB RTX6000 GPU. For the 3D generation process, the SMPL-X fitting was performed with 3000 iterations in 3 minutes, followed by skin color inpainting on SMPL-X Gaussians for 100 iterations in 30 seconds. Reconstruction and disentanglement optimization required 3000 iterations, completed in 12 minutes. In video reconstruction, SMPL-X fitting aligned 14 frames in 6 minutes for in-the-wild videos. The 4D-Dress [236] experiments involved 1000 iterations for clothing deformation over 18 minutes.

TABLE 5.2: **User study rates quality of generated 3D Gaussians from 1-5. The higher the better.**

Metric	Image Consistency ↑	Overall Quality ↑
DreamGaussian	2.017	1.852
LGM	2.338	2.017
Disco4D	**3.142**	**3.037**

5.3.2 3D Generation

Generation and Disentanglement. Our generation and disentanglement results are presented in Figure 5.3 and Table 5.1. We assessed the disentanglement quality using the Synbody [237] and CloSe [238] datasets, rendering 30 and 110 clothed human meshes respectively from four angles and evaluating CLIP-similarity, PSNR, SSIM, and LPIPS for various poses and views within the CloSe dataset. Disco4D leverages diffusion models without requiring training on human specific datasets. Therefore, we compare it with DreamGaussian [115] and LGM [116] which reconstruct 3D objects from diffusion models. Additionally, we conducted comparisons with SHERF, a human-centric baseline for evaluating novel poses and views. Figure 5.3 shows Disco4D has higher fidelity and better geometry for body parts such as face and limbs due to the representation using SMPL-X Gaussians. It outperforms DreamGaussian and SHERF on SynBody and CloSe benchmarks. Disco4D performs worse than LGM on novel views, likely due to its optimization of Gaussians in canonical space for pose generalization, compromising view-specific detail.

Editing. We can edit specific clothing appearance given an image or text prompt, repose the person and transfer person characteristics. The disentanglement allows fine-grained editing and modification of individual assets without affecting other assets, and stacking multiple edits (Figure 5.3).

User study. We conducted a user study to evaluate the generative quality of our image-to-3D Gaussians reconstruction on random in-the-wild images from SHHQ, detailed in Table 5.2. This study focuses on reference view consistency and overall generation quality, crucial aspects in image reconstruction tasks. We rendered 360-degree rotation videos for 25 images generated by DreamGaussian, LGM, and Disco4D. We invited 43 volunteers to rate 24~27 mixed samples from these methods on image consistency and overall model quality, yielding 1080 valid scores. As

FIGURE 5.4: **Qualitative evaluation on ITW images.**

FIGURE 5.5: **Qualitative evaluation on avatars clothed in dress.**

shown in Table 5.2, Disco4D was preferred, demonstrating better alignment with the original image content and superior overall quality.

In-the-wild evaluation. Our focus on studio and synthetic datasets (e.g., Synbody, CloSe, and 4DDress) was due to the availability of ground-truth data from multiple views, enabling rigorous quantitative evaluation. ITW images lack such ground-truth data, making comparisons challenging. Nevertheless, our solution applies to ITW images, with some examples shown in Fig. 5.4. Examples of avatars clothed in dress are added in Fig 5.5, driven with poses from subjects in Fig. 5.4.

Takeaway. Anchoring clothing Gaussians to a fixed SMPL-X body yields sharper face and limb geometry than generic object-level initialisations (e.g., DreamGaussian, LGM). Separately, representing clothing as disentangled Gaussian layers supports fine-grained, asset-level editing without affecting adjacent components.

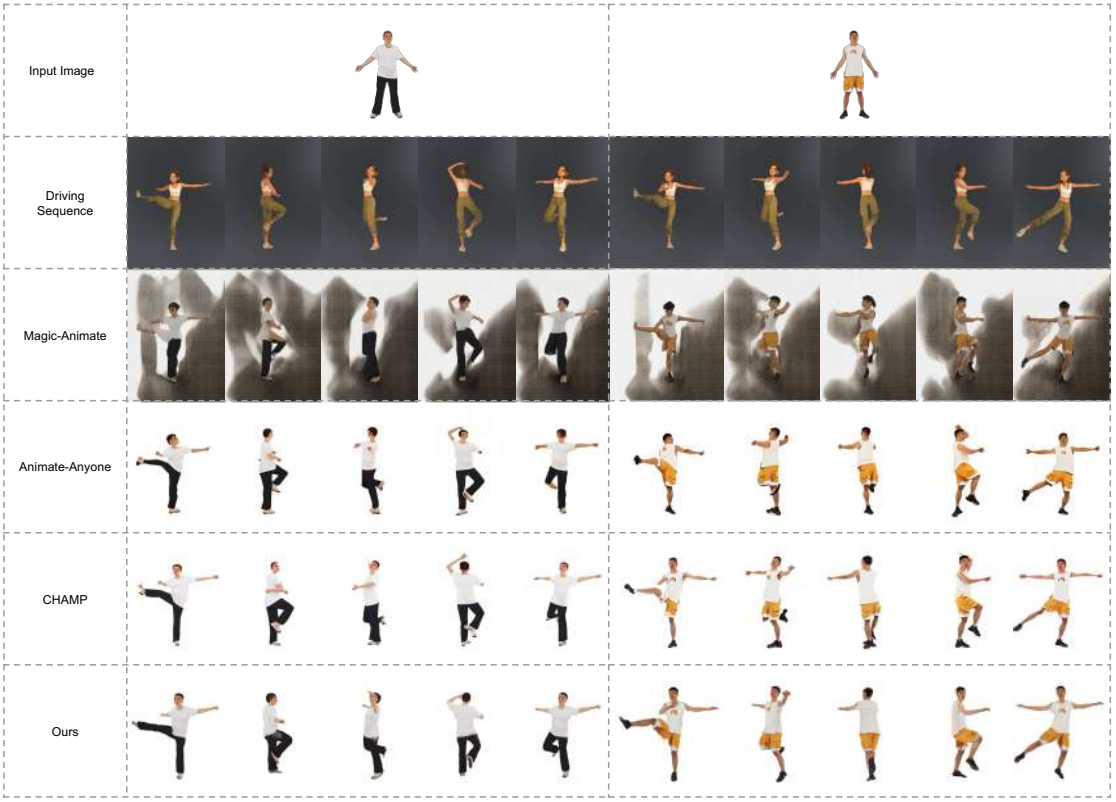


FIGURE 5.6: **Comparison to 2D animation methods.** Compared to Magic-Animate and Animate-Anyone, we have better preservation of body shape and details. Compared to CHAMP, we have better geometry and consistency.

5.3.3 4D Animation

Pose-Driven Animation. Disco4D generates canonical Gaussians that can be animated with any pose sequence. Figure 5.6 demonstrates our animation capabilities and compares them with current SOTA 2D animation methods. Using identical inputs—a single frame and pose sequence—our approach more effectively preserves the body shape and fine details such as facial features and clothing. It surpasses Animate-Anyone [28] and Magic-Animate [29] in accurately modeling fine-grained body parts like hands and faces, and exhibits greater consistency compared to CHAMP [32]. The disentanglement feature of Disco4D further allows for direct manipulation of Clothing Gaussians, as shown in Figure 5.9.

4D Reconstruction. For the 4D-Dress Dataset [236], we evaluated 8 sequences, assessing CLIP similarity scores against ground-truth meshes and disentangled assets, along with novel view performance (PSNR, SSIM, LPIPS) from four viewpoints. Table 5.3 summarizes our quantitative results, benchmarking Disco4D against

TABLE 5.3: **CLIP-embedding loss for generated humans and segmented assets, and performance (PSNR, SSIM, LPIPS) comparison on 4D-Dress across various video-to-4D methods.**

	All ↑	Assets ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DreamGaussian4D	0.784	0.769	20.54	0.93	0.080
MonoHuman	0.762	0.743	20.22	0.92	0.086
GART	0.800	0.772	18.81	0.92	0.086
GaussianAvatar	0.822	0.768	20.01	0.93	0.069
DreamGaussian4D (LGM init)	0.809	0.795	19.16	0.93	0.086
DreamGaussian4D (Disco4D init)	0.870	0.849	21.02	0.93	0.065
Disco4D (reposed)	0.853	0.774	23.94	0.95	0.049
Disco4D (reposed)+learned deformations	**0.900**	**0.865**	**25.46**	**0.96**	**0.035**

FIGURE 5.7: **Qualitative comparison of 4D generation between DreamGaussian4D, MonoHuman, GART, GaussianAvatar, and Disco4D.**

existing video-to-4D general GS approaches, such as DreamGaussian4D [173], as well as human-centric GS methods, including MonoHuman [177], GART [180], and GaussianAvatar [179]. We evaluate on monocular videos comprising 14 frames, captured from a limited front-view perspective, without full-body visibility across frames.

Disco4D outperforms MonoHuman [177], GART [180], and GaussianAvatar [179] (Table 5.3) as these methods reconstruct using known video information, unable to model unseen regions. Consequently, these methods cannot accurately model

FIGURE 5.8: **4D reconstruction results on 4D-Dress Dataset.**

back views from front-facing videos, leading to artifacts in other perspectives and canonical space (see Figure 5.7). In contrast, Disco4D first performs reconstruction and subsequently incorporates details, such as clothing deformation, from the input frames, enabling consistent reconstruction even in unseen viewpoints.

While DreamGaussian4D [173] is capable of modeling back-view information, the details remain coarse. Our results demonstrate that initializing with our model from the first frame (DreamGaussian4D Disco4D-init) significantly outperforms other initialization methods (DreamGaussian4D-LGM init, DreamGaussian init) in both fidelity and geometry (Table 5.3). Nevertheless, without incorporating human priors, DreamGaussian4D [173] still faces challenges, such as missing limbs and difficulty modeling fine details like facial features (see Figure 5.8.

Reposing our canonical avatar enables us to align the body and assets accurately with the inferred postures from the source video, yielding high-quality reconstruction of faces, hands, and garments. Our reposed method surpasses DreamGaussian4D in geometry and fidelity by incorporating human priors. However, reposing alone cannot capture clothing dynamics. To address this, our disentangled approach models clothing deformations on the reposed Gaussians, guided by a diffusion model. As demonstrated in Figure 5.8 and Table 5.3, this process enhances the accuracy of clothing resemblance to the ground truth. The combination of asset repositioning and learned deformations improves modeling quality, with repositioning handling

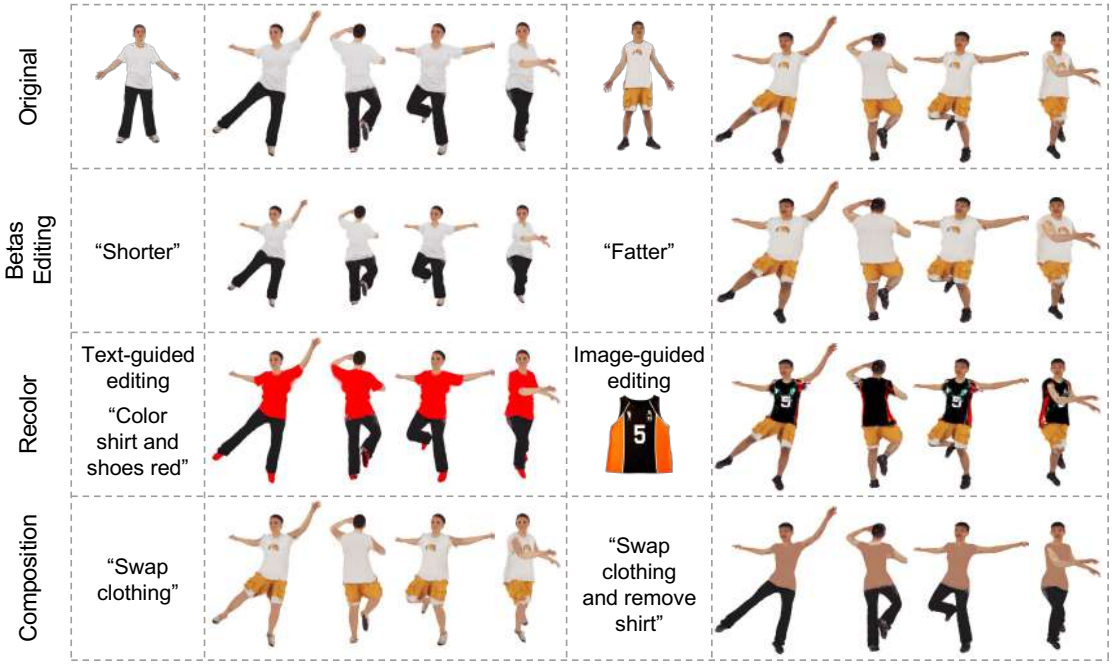


FIGURE 5.9: **First frame Editing and Animation**. Betas Editing, Recoloring (Text/Image-guided), Composition (Removal, Swap).

pose-driven changes and learned deformations simulating dynamic asset movements as observed in the driving video.

Takeaway. Decoupling body deformation (via SMPL-X reposing) from clothing deformation (via a diffusion-guided residual) enables temporally coherent 4D reconstruction from limited views. Reposing accounts for pose changes, while the residual captures asset-dependent motion.

4D Editing. For a normal pipeline in character animation, editing the person in the video requires high consistency throughout all frames. For pose-driven animation methods, first frame editing and generation is required. Our method directly edits the Gaussians, which is more straightforward, fine-grained and consistent. This is seen from Figure 5.9. Extended visualizations and results showcasing 3D generation and disentanglement, pose-driven animation, video-to-4D reconstruction, and fine-grained editing of animated outputs are demonstrated in the accompanying demo results in our project page at https://disco-4d.github.io/.

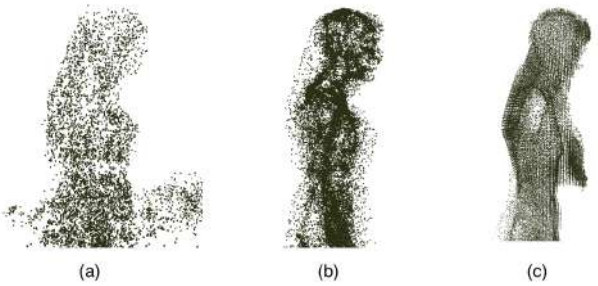


FIGURE 5.10: **Ablation of initialization.** (a) Random Initialization (b) SMPL-X Initialization (c) Visual Hull Initialization.

FIGURE 5.11: **Ablation of points geometry (left) and editing results (right). Points ("All") are visualised with a Gaussian Scale of 0.1.**

5.3.4 Ablation Studies

Initialization of Clothing Gaussians. This process is crucial for high fidelity reconstruction. As shown in Figure 5.10, we evaluate different strategies, including random, surface, and hull-based initialization. Hull-based initialization significantly enhances the model accuracy and realism over other methods. Initialization directly on the SMPL-X surface often leads to inaccurate geometries, particularly with complex or loose garments, creating elongated, thin Gaussians and visual artifacts. In contrast, hull-based initialization captures garment details more effectively and maintains pose consistency, closely aligning with the true geometry of the clothed body.

Geometry of Clothing Gaussians. Figure 5.11 highlights the differences in clothing geometry between DreamGaussian [115], LGM [116] and Disco4D. In DreamGaussian, all points are confined within the body geometry, whereas in LGM, about half of the points extend beyond the SMPL-X body. Removing internal points leaves sparse, translucent representations for clothing. This sparsity suggests

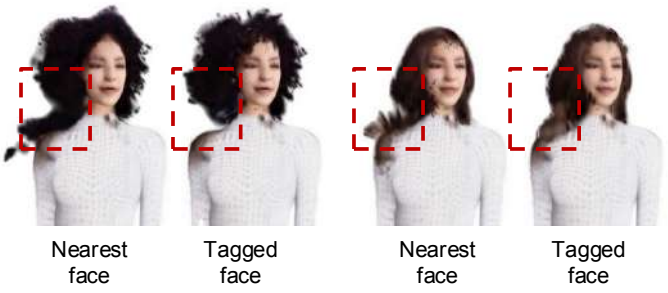


FIGURE 5.12: **Visualization of hair tagging.**

reliance on internal points for visual representation, failing to accurately depict the object's geometry where appearance should primarily originate from surface points. Often, clothing Gaussian points are incorrectly positioned inside the body's hull rather than on the surface. To better represent clothing geometry, Disco4D positions all clothing Gaussians externally to the SMPL-X body mesh, accurately reflecting the garment's actual physical characteristics.

Clothing editing. Figure 5.11 shows our editing results with the prompt "Color the top pink". Disco4D allows for precise editing of the targeted clothing without affecting other areas.

Hair tagging. In our approach, hair Gaussians are tagged to head faces rather than the nearest face during reposing. Reposing hair Gaussians according to the nearest face, as commonly done in previous works, often results in artifacts such as disjointed hair (Figure 5.12). By leveraging the learned identity encoding, we assign a unified identity to hair Gaussians, enabling them to be reposed cohesively as a single entity, thereby preserving the structural integrity of the hair during transformations.

5.4 Discussion

Despite achieving impressive results, some failure cases still exist, as shown in Figure 5.13. Its performance depends on robust, pixel-aligned SMPL-X estimation, which remains unsolved for challenging poses (Figure 5.13a). It occasionally fails for poor visual hull initialization, which is common in difficult poses (Figure 5.13b). Lastly, misclassification by the segmentation model can lead to poor disentanglement, such as arms being misclassified as "top" clothing (Figure 5.13c).

FIGURE 5.13: **Failure cases of Disco4D.** (a) Poor SMPL-X estimation (b) Poor visual hull initialization (c) Misclassification of clothing categories.

The extraction of mesh assets from clothing Gaussians using Local Density Query, as per DreamGaussian [115], currently loses fine-grained details. Enhancing the detail level of geometry derived from clothing Gaussians could bolster the utility of reconstructed assets in animation and simulation applications. Furthermore, the initialized visual hulls obtained from multi-view SMPL-X guided images are often of suboptimal quality and suffer from poor side and back views, necessitating refinement. Improving pose guidance models to achieve more accurate visual hulls could alleviate the need for extensive refinement. In addition, future works could look into modeling multi-layered clothing and reconstructing the occluded clothing.

Loose clothing and physical dynamics. However, the dynamics are geometric rather than physics-based. Clothing deforms with the body via learned offsets, without modelling inertia, collisions, or material properties. As a result, while appearance is well reconstructed from the input and diffusion-guided views, motion under novel poses follows the body and lacks physically consistent secondary dynamics. In particular, loose clothing Gaussians that track large body deformations can appear "broken" under extreme poses. Two promising directions are: (i) integrating a lightweight physics model to enable physically plausible garment motion; and (ii) representing clothing as an explicit surface (e.g., a mesh) to better preserve structural coherence during deformation.

Societal considerations. While Disco4D enables applications in content creation and entertainment, it also raises risks such as deepfake misuse and IP concerns, underscoring the need for appropriate safeguards and regulation.

5.5 Conclusion

We propose Disco4D, a novel approach for the generation of 3D animatable clothed human Gaussians from a single image, emphasizing high-fidelity detail and separation of assets. We manage to compositionally generate separate components, such as haircut, accessories, and decoupled outfits. Our core insight is the fixing of SMPL-X Gaussians, fitting segmented Gaussians over SMPL-X Gaussians, and application of diffusion models to enhance 3D reconstruction, including modeling occluded parts not visible in the input image. Its capability to separate assets offers significant advantages, including localized, fine-grained editing of individual assets and enhanced animatability.

Within the broader context of this thesis (Fig. 1.1), Disco4D extends robust body mesh recovery to clothed and disentangled avatar modeling, enabling structured representations that support editing and animation. At the same time, achieving consistent and high-fidelity asset recovery remains dependent on reliable canonical multi-view supervision, which is addressed in Chapter 6.

Chapter 6

Using Video Diffusion Models to Generate Animatable and Customizable Human Avatars with Reusable Assets

6.1 Introduction

This chapter addresses the fourth and final stage of the thesis pipeline shown in Fig. 1.1: identity-level controllable generation on top of the layered avatar representation introduced in Chapter 5. A residual limitation of the previous stage motivates the approach developed in this chapter: the absence of canonical multi-view supervision needed for consistent asset recovery.

Developing high-fidelity 3D human avatars that are both **animatable** and **customizable** from a single image is critical for applications like animation, virtual try-on, and interactive environments. Additionally, achieving disentanglement, where individual attributes (e.g., skin texture, clothing, accessories) can be independently modified, is crucial for enabling flexible customization without regenerating the entire avatar.

Unfortunately, significant challenges remain. Despite advances in single-image human reconstruction [19, 20, 24, 239], animating these models remains difficult.

They are typically aligned to non-canonical input poses, requiring complex rigging for motion control. Existing methods for clothed human reconstruction [18–27] fuse body and clothing into a single-layer, non-animatable mesh. This prevents effective separation of clothing, hair, and accessories, especially in self-contact regions. As a result, these models are unsuitable for tasks requiring dynamic customization and layered representations.

A promising approach is to use **multi-view canonical pose images**, which provide viewpoint consistency and structural integrity for animation. A standardized pose, such as A-pose or T-pose, offers a stable baseline for motion retargeting and reduces articulation artifacts. Multi-view images also facilitate disentangled asset generation and enable independent reconstruction of clothing, hair, and accessories. However, capturing them in practice requires complex camera setups, precise calibration, and controlled environments. Layered reconstruction methods [144, 145, 240, 241] attempt to address this but often rely on user-provided rotating videos, limiting scalability and practicality.

Inspired by the success of diffusion models in multi-view object generation [105, 111], some works extend these approaches to multi-view human image generation [146, 147], but focus only on static reconstruction. Other diffusion-based human animation methods [28, 30–32] enable pose-controlled synthesis but suffer from view inconsistencies, motion artifacts, and poor preservation of fine shape, facial details, and clothing. These limitations stem from their reliance on 2D video training. Recent works such as CharacterGen [33] and EN3D [34] synthesize multi-view canonical images for riggable 3D avatar reconstruction. However, they still struggle with fine-detail preservation and offer limited control over individual assets.

Relation to Disco4D (Chapter 5). Among recent approaches, Disco4D is the closest prior work, reconstructing a disentangled clothed human from a single image using SMPL-X and layered Gaussians. However, it has three limitations that motivate a new approach rather than an incremental extension. First, it requires a near-canonical input pose, whereas `ReposeHuman` accepts arbitrary poses and reposes the subject into a canonical A-pose via a trained video diffusion model. Second, its multi-view supervision comes from existing pose guidance models with low-quality side and back views, whereas `ReposeHuman` trains a dedicated video diffusion model with dense pose conditioning. Third, its Gaussian-to-mesh extraction loses fine

detail, whereas `ReposeHuman` recovers clean asset-level meshes via 2D Gaussian Splatting with normal supervision.

Building on this analysis, we propose `ReposeHuman`, a 3D clothed human generation method that separates body and clothing while ensuring structural consistency from a single image. Our key contributions are:

- **A video diffusion model for dense multi-view image generation in arbitrary poses.** The model supports diverse body shapes and sizes, and produces riggable avatars even from partial views. To improve identity and pose control, we build on two existing conditioning mechanisms—IP-Adapters [35] and LoRA [36]—which we compose into task-specific modules: Face IPA, Clothing IPA, and Pose LoRA.
- **A dataset of 3D disentangled human assets** for independent asset control and structured pose conditioning.
- **An improved Gaussian Splatting pipeline for geometrically accurate, high-fidelity asset recovery.** We combine 2D Gaussian Splatting, Gaussian Grouping, and normal supervision to reconstruct smooth and clean clothing meshes from disentangled Gaussians.

We demonstrate that `ReposeHuman` generates high-quality, customizable avatars suitable for animation and virtual environments. It provides a scalable multi-view generation solution from a single image, ensuring pose consistency and asset editing. By addressing animation fidelity and disentanglement, `ReposeHuman` improves realism and adaptability for interactive and immersive applications.

6.2 Preliminary

2D Gaussian Splatting (2DGS). 2DGS [242] represents surfaces using planar Gaussian disks tightly aligned to surface geometry. Each 2D Gaussian is defined by a center point μ, a covariance matrix Σ, an opacity α, and a view-dependent color c parameterized by spherical harmonic coefficients f.

Each disk is defined in a local tangent frame by orthogonal vectors t_u and t_v, with scale parameters (s_u, s_v) controlling elliptical variance: $P(u, v) = \mu + s_u t_u u + s_v t_v v$.

The Gaussian at (u, v) is: $G(u, v) = e^{-\frac{1}{2}(u^2+v^2)}$. During rendering, an explicit ray-splat intersection computes the contribution of each Gaussian, eliminating perspective errors found in 3DGS [114] projections. Pixels are accumulated by front-to-back alpha blending: $C = \sum_i T_i \alpha_i c_i$, with $T_i = \prod_{j=1}^{i}(1 - \alpha_j)$. Unlike 3DGS [114], which uses volumetric Gaussians, 2DGS directly defines surface-aligned disks for improved geometric accuracy and multi-view consistency.

SMPL-X parameterization. SMPL-X [5] extends SMPL [1] incorporating detailed face and hand deformations to capture expressive motions. It expands the SMPL joint set with additional joints for facial features, toes and fingers. SMPL-X is defined as: $M(\beta, \theta, \psi) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{3N}$, where $\theta \in \mathbb{R}^{3K}$ represents the pose (with K being the number of body joints), $\beta \in \mathbb{R}^{|\beta|}$ models body shape, and $\psi \in \mathbb{R}^{|\psi|}$ captures facial expressions. Further details are in [5].

Video Diffusion Models. These models generate video sequences by iteratively denoising noise-initialized latent representations. Traditional UNet architectures model spatiotemporal latents but struggle with long-range dependencies and temporal consistency. Diffusion Transformer (DiT) architectures improve performance by capturing long-range temporal dependencies and enhancing frame-to-frame coherence.

6.2.1 Video Diffusion Components

I2V Video diffusion models. For our denoising model, we opt for the open-source Image-to-Video (I2V) CogVideoX [243] diffusion model that employs a diffusion transformer architecture. It utilizes a 3D Variational Autoencoder (VAE) to project videos from the pixel domain into a latent space, where the latents are then patchified and unfolded into a long sequence denoted as z_{noise}. Simultaneously, textual input is encoded into text embeddings z_{text} using T5[244]. In the I2V models, image is an additional condition alongside the text z_{image}. The image is passed through 3D VAE and concatenated with the noised input in the channel dimension. Subsequently, z_{text}, z_{noise} and z_{image} are concatenated along the sequence dimension. The concatenated embeddings are then processed through a stack of expert transformer blocks, which integrates Adaptive Layer Normalization for better alignment and 3D Rotary Positional Encoding (RoPE)[245] to enhance the model's ability to capture temporal dynamics and long-range dependencies in video frames.

Finally, the model output is unpatchified to restore the original latent shape and subsequently decoded using a 3D causal VAE decoder to reconstruct the video.

Low-rank Adaptation (LoRA) [36] is a parameter-efficient fine-tuning technique that reduces the computational cost of adapting large-scale pre-trained models by injecting low-rank decomposition matrices into their weight updates. LoRA [36] targets the residual component of the model, denoted as ΔW, which is added to the original weight matrix, yielding the updated weights as:

$$W' = W + \Delta W.$$

In this formulation, ΔW is expressed as the product of two low-rank matrices:

$$\Delta W = AB^T,$$

where $A \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{m \times d}$, and $d < n, d < m$. By focusing on the smaller low-rank matrices A and B, rather than the full-weight matrix W, LoRA effectively reduces both computational and memory costs during the training process.

Image Prompt Adapters (IP-Adapter) [35] was initially introduced as a lightweight plug-and-play module to enable image prompt conditioning and integrate identity-preserving features in pretrained text-to-image diffusion models, enabling efficient adaptation with minimal computational overhead. We extend this approach by integrating IP-Adapter into image-to-video (I2V) diffusion models, significantly improving identity preservation, particularly in clothing and facial consistency.

The proposed IP-Adapter consists of two key components: (1) an image encoder that extracts visual features from an image prompt and (2) adapted cross-attention modules that inject these features into the pretrained I2V video diffusion model using decoupled cross-attention. We employ Sapiens [246] as the image encoder to enhance feature extraction.

Given query features $\mathbf{Z}$ and text features c_t, the standard cross-attention mechanism is formulated as $\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}$, where $\mathbf{Q} = \mathbf{Z}W_q$, $\mathbf{K} = c_t W_k$, and $\mathbf{V} = c_t W_v$ are the query, key, and value matrices. To incorporate image features c_i, a new cross-attention layer is introduced, yielding: $\mathbf{Z}'' = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}'^\top}{\sqrt{d}})\mathbf{V}'$, where $\mathbf{K}' = c_i W'_k$, $\mathbf{V}' = c_i W'_v$. The final formulation combines text-based and image-based attention, ensuring both textual and visual

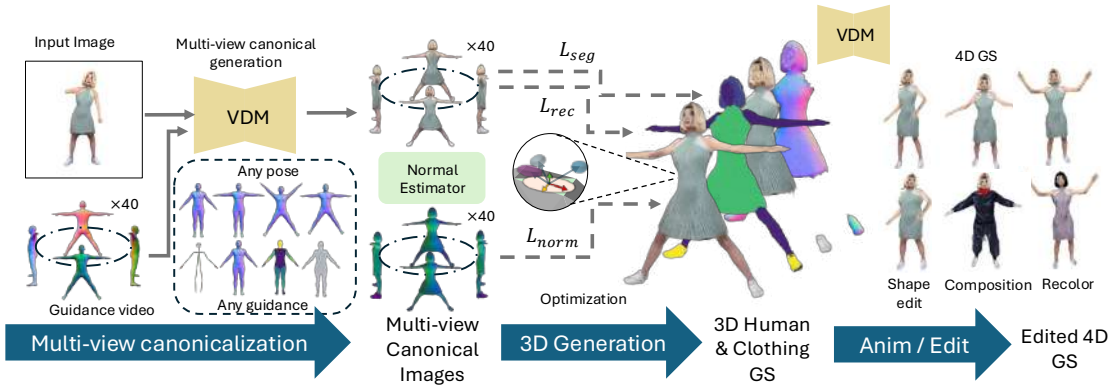


FIGURE 6.1: **Framework Overview of `ReposeHuman`.**

conditioning: $\mathbf{Z}^{\text{new}} = \text{Softmax}(\frac{\mathbf{QK}^\top}{\sqrt{d}})\mathbf{V} + \text{Softmax}(\frac{\mathbf{QK'}^\top}{\sqrt{d}})\mathbf{V'}$. Notably, the original transformer blocks remain frozen, with only the newly introduced IP-Attention weights W'_k and W'_v being trainable. This ensures an efficient and lightweight integration of image features, allowing seamless identity preservation in video diffusion models.

6.3 Methodology

Given a single image or text prompt, we aim to generate an animatable, disentangled 3D GS avatar. The avatar is created across multiple canonical poses and rigged with an aligned SMPL-X model. The disentangled assets allow for flexible editing and composition.

An overview of our `ReposeHuman` framework is provided in Figure 6.1. It has three stages: (1) generate dense multi-view canonical images from an input image or text with pose guidance using a reposing model, (2) reconstruct a disentangled 3D GS avatar, and (3) apply optional editing or animation, including clothing deformation.

6.3.1 Reposing Model Architecture

To generate multi-view images, we design a reposing model upon a Video Diffusion Transformer, as shown in Figure 6.2. We adopt the open-source Image-to-Video (I2V) model, CogVideoX [243]. It processes textual input via T5 [244] embeddings, concatenating image conditions with noisy latents in the sequence dimension for

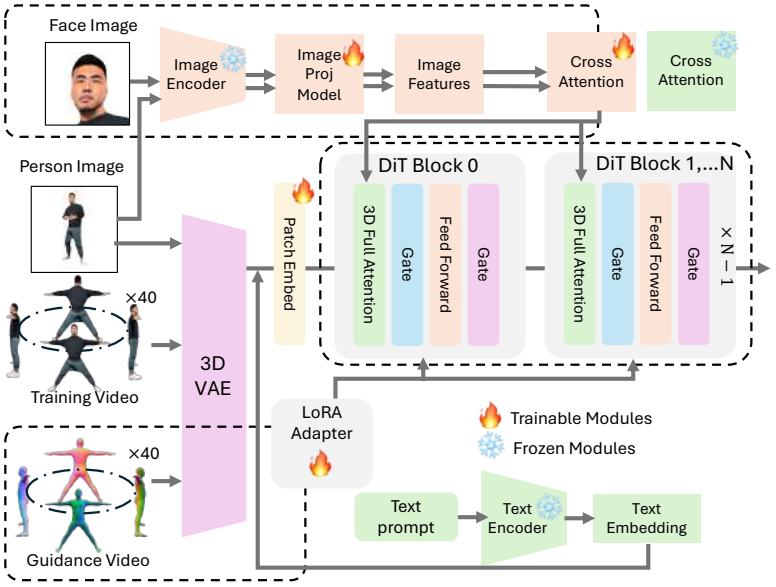


FIGURE 6.2: **Our video diffusion model generates multi-view images with improved identity consistency using person and face IPAs. A LoRA adapter enables control over pose and shape.**

improved alignment. Additionally, Adaptive Layer Normalization and 3D Rotary Positional Encoding (RoPE) [245] enhance temporal coherence and long-range dependencies in generated video frames.

During denoising, the reference image is encoded into VAE features $F_i \in \mathbb{R}^{B\times1\times C\times HW}$, where B is batch size, C is channels, and HW is spatial tokens. Video noise latents are $F_v \in \mathbb{R}^{B\times f\times C\times HW}$, where f is the number of output frames. To maintain identity consistency, reference image features are fused with latent features, enabling self-attention-based feature interaction.

Our approach introduces two lightweight yet effective enhancements to the I2V model:

1) Dual IP-Adapter Integration. We add two Image Prompt (IP) Adapters [35] to enhance identity consistency: a Person IPA for body and clothing details, and a Face IPA for fine facial features. Both inject reference image features into cross-attention layers. We use Sapiens [246] as the image encoder. Only the IP-Adapter weights are trained, keeping the base model frozen. This improves identity retention and generation quality across views.

2) Control Signal Integration. We guide generation using structured control signals, including keypoints, normals, depth, and semantic maps. Pose control is

FIGURE 6.3: **Some examples of our `ReposeHuman` dataset.**

applied using Low-Rank Adaptation (LoRA)[36], which injects low-rank decomposition matrices into weight updates with minimal overhead (see Section 6.2.1). Our model supports multiple canonical poses, such as X-pose, DA-pose, and T-pose, improving generalization across body types, clothing styles, and hairstyles. We also inject dense SMPL-X [5] signals, following [32], to better preserve body shape. This avoids distortions seen in EN3D [34] and CharacterGen [33], which rely only on sparse keypoints. The rendered pose guidance video is encoded by a 3D VAE into pose latents. These are concatenated with video noise latents before each denoising step. This lets the model jointly process image conditions and pose guidance during generation.

6.3.2 Disentangled 3D Synthetic Dataset

Existing datasets face key limitations: (1) 3D static datasets like THuman [150], 4D-Dress [236], and CloSe [238] lack pose and body shape diversity, restricting generalization. (2) Riggable datasets such as EN3D [34] and RenderPeople are small-scale and lack SMPL-X ground truth, complicating dense body alignment. Large-scale datasets like BEDLAM [247] and SynBody [237] do not release their 3D models and assets. (3) All existing datasets have simplistic clothing textures, limiting real-world applicability.

To address these gaps, we build a high-quality synthetic disentangled 3D dataset with diverse body shapes, detailed clothing textures, and expanded pose variations. We integrate open-source assets to enhance realism and diversity:

1. **SMPL-X textures:** High-quality facial textures from FFHQ-UV [248] are transferred to FLAME. SMPL-X body textures are sampled from BEDLAM [247] and color-corrected to match faces.

TABLE 6.1: **Comparisons of our dataset with existing ones.**

Dataset	IDs	SMPL-X	Animatable	Layered
THuman	200	✓	✗	✗
THuman2.1	2500	✓	✗	✗
2K2K	2050	✗	✗	✗
EN3D	900	✗	✓	✗
4D-Dress	64	✓	✗	✗
CloSe	3000	✓	✗	✗
`ReposeHuman`	50K	✓	✓	✓

2. **Garment and texture diversity:** GarmentCode [249] clothing assets are adapted from their custom body model to SMPL-X. A text-conditioned texture generation model enhances texture variations.
3. **Hair modeling:** HAAR [250] hair assets are mapped from their custom head model to SMPL-X. Hair textures are synthesized from scalp textures for realism.

The dataset pipeline involves: (1) generating SMPL-X bodies with varied shapes and poses, (2) mapping clothing and hair assets, and (3) synthesizing high-resolution body and clothing textures. All samples are consistently aligned and reposable with SMPL-X parameters.

Figure 6.3 shows dataset samples. Table 6.1 compares our dataset to existing ones in scale, diversity, and consistency. Our dataset contains over 50K identities and 50K distinct poses, covering representation across geographical origins, clothing styles, body shapes, age groups, and genders. Each sample includes: (1) a reposable 3D textured human model with clothing and hair, along with ground-truth SMPL-X parameters; (2) dense multi-view renders in canonical and posed states.

For evaluation, we set aside 50 canonicalized samples from CloSe [238] samples and 8 video samples from 4D-Dress[236] for 3D generation and 4D animation experiments.

6.3.3 Training Strategy

Robust augmentation. Input images are often segmented and may contain imperfections. To improve generalization, we apply robust augmentations: (1) half-body cropping for partial views, (2) object-shaped cropping and pasting using OccludedPASCAL3D+[251] to simulate imperfect segmentation, and (3) random scaling and translation to prevent over-reliance on centered subjects. These augmentations increase model resilience to input variations.

Two-stage training. We first train the model on a large set of 3D models with keypoint-conditioned supervision. In the second stage, we introduce dense body shape controls using rendered SMPL-X data, which requires models with accurate SMPL-X fitting. This joint training ensures close alignment between generated images and pose guidance. We also apply extra facial consistency losses and augmentations to further enhance identity preservation and robustness.

6.3.4 Generation, Animation and Editing of Disentangled GS Avatar

3D Disentangled Generation. The consistency of multi-view images in canonical space makes avatar generation efficient. Inspired by Disco4D [228], we model Gaussians as separable layers on the human body, introducing key modules for better efficiency and fidelity.

Given an image, we estimate SMPL-X parameters with an off-the-shelf model [229]. The body is posed to a canonical or target pose. We render 40-view pose guidance videos and apply our reposing model to generate dense, high-quality multi-view images. Body Gaussians, colored to match skin, are bound to each SMPL-X triangle to form S_{body} under skin color consistency loss. Clothing Gaussians S_{cloth} are optimized separately and disentangled from S_{body}. We enforce spatial separation using SDF loss and pruning. Identity encoding [234] and multi-view segmentation guide grouping and densification of S_{cloth} for fine detail recovery. The result is a high-quality avatar with structured body and clothing layers, enabling animation and editing.

We introduce several improvements over Disco4D:

- **High-fidelity multi-view generation:** Our reposing model directly generates high-fidelity multi-view images in canonical or target poses. This eliminates the need for visual hull or LGM-based Gaussian initialization and minimizes refinement.
- **SMPL-X aligned generation:** Our pipeline eliminates explicit SMPL-X fitting by leveraging reposed multi-view generation, reducing computational cost versus Disco4D.

- **Geometrically accurate 2D Gaussians:** We incorporate 2D Gaussian Splatting (2DGS)[242], which uses planar Gaussian disks to achieve better view consistency and smoother geometry than 3DGS.
- **Clothing mesh-guided embedding:** We recover a clothing mesh and embed Gaussians on its surface for better alignment and stability under deformation, especially for skirts and dresses, which often tear under large motions in Disco4D.

4D animation Like Disco4D, `ReposeHuman` animates avatars by (1) driving SMPL-X sequences from motion datasets or 2D video estimates, (2) learning fine clothing dynamics from monocular videos while body Gaussians remain fixed, and (3) independently animating clothing Gaussians attached to the recovered clothing mesh.

We use a deformation network D_N to predict position, rotation, and scale changes of clothing Gaussians S' at time t, yielding $S'' = D_N(S', t)$, following DreamGaussian4D [173]. Unlike [173], which deforms all Gaussians, we only deform clothing Gaussians while SMPL-X drives body motion. The network starts from zero deformation for stability and uses skip connections to maintain smooth gradients. Body Gaussian color and opacity remain fixed to ensure temporal consistency despite per-frame diffusion.

Instead of using Zero-1-to-3-XL [106] as in Disco4D to hallucinate unseen motions with SDS loss, our reposing model directly generates unseen views, enabling stronger supervision through reconstruction loss and yielding higher-fidelity results.

4D editing. We enable fine-grained editing by isolating and modifying only Gaussians of the target object, leaving others fixed. Object removal is done by deleting the corresponding Gaussians. Color edits use in-painting or text guidance to modify only the color (SH) parameters, preserving geometry.

6.3.5 Experiment Setup

The 2D generation experiments were trained on 8×H100 GPUs for 80k steps over 5 days for both stages. The 3D generation experiments used a single 24GB RTX3090 GPU, while the 4D experiments utilized a 40GB RTXA6000 GPU. For 3D generation, SMPL-X optimization ran for 1k steps in 30s, followed by skin color inpainting on SMPL-X Gaussians for 100 iterations in 30s. Generation and disentanglement

TABLE 6.2: **Quantitative results of multi-view canonical image generation on the CloSe.**

	SSIM ↑	LPIPS ↓	PSNR ↑	CLIP ↑	Asset CLIP ↑
CharacterGen [33]	0.8160	0.2057	11.3275	0.7597	0.6948
AnimateAnyone [28]	0.8867	0.1141	15.1646	0.8480	0.8047
MimicMotion [31]	0.8824	0.1253	14.1805	0.8163	0.7722
CHAMP [32]	0.9205	0.0717	17.7750	0.9157	0.8672
MusePose [32]	0.9026	0.0915	16.1468	0.8887	0.8477
`ReposeHuman`	**0.9221**	0.0625	**18.8452**	**0.9263**	**0.8890**
`ReposeHuman` w/o dataset	0.8986	0.1053	15.7912	0.8923	0.8421
`ReposeHuman` w/o augmentation	0.9208	**0.0619**	18.8331	0.9245	0.8560
`ReposeHuman` w/o dense controls	0.9132	0.0815	17.7782	0.9053	0.8601

optimization required 3000 iterations, completed in 6 minutes—half the time of Disco4D [228]. In 4D-Dress [236] video reconstruction, clothing deformation was optimized over 1000 iterations in 15 minutes.

6.4 Evaluation

6.4.1 2D Canonical Human Generation

We evaluate `ReposeHuman` against leading methods: CharacterGen [33], AnimateAnyone [28], MimicMotion [31], CHAMP [32], and MusePose [30]. CharacterGen generates only four A-pose views, so we limit its evaluation accordingly. The other methods are compared over 40 canonical views rendered from CloSe [238] meshes.

Table 6.2 and Figure 6.4 show quantitative and qualitative results. The CloSe dataset includes 50 clothed human meshes rendered at 40 viewpoints for full 360-degree coverage. We report LPIPS, SSIM, PSNR, CLIP similarity, and Asset CLIP (on segmented clothing regions).

`ReposeHuman` outperforms all baselines across metrics. It achieves the highest SSIM (↑), PSNR (↑), and CLIP scores (↑), and the lowest LPIPS (↓), as shown in Table 6.2. Compared to methods trained on 2D video data [28, 31, 32], `ReposeHuman` benefits from our 3D disentangled dataset and dense pose controls. We show that training only on 3D data can outperform 2D-based approaches in reposing to new poses and achieving superior multi-view consistency. Ablations confirm this: removing

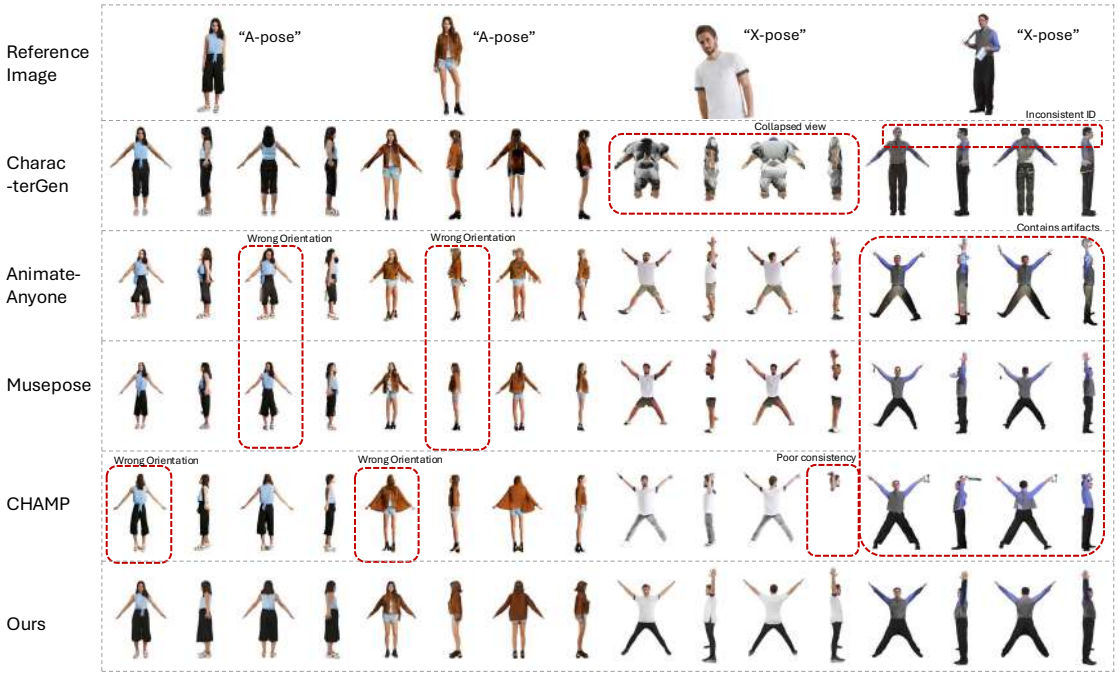


FIGURE 6.4: **Qualitative comparisons of multi-view canonical video generation on in-the-wild images from the SHHQ dataset.**

the dataset (`ReposeHuman` w/o dataset) significantly reduces all scores, highlighting the importance of 3D supervision. Removing augmentation (`ReposeHuman` w/o augmentation) yields minimal drop on average metrics, but Figure 6.5b shows that robust augmentation helps especially on difficult samples. Removing dense controls (`ReposeHuman` w/o dense controls) lowers performance, confirming the value of strong conditioning. CHAMP [32], which also uses dense conditioning, ranks second behind `ReposeHuman`.

Figure 6.4 illustrates that `ReposeHuman` produces superior consistency and robustness. It successfully disentangles the subject from artifacts, even under half-body crops and imperfect segmentation. AnimateAnyone [28] and MusePose [30] often produce incorrect back and side views, likely due to front-view biases in their training. CharacterGen [33] is limited to A-pose and shows poor identity retention with partial crops. Compared to CHAMP [32], our method maintains better body shape consistency across views. Overall, `ReposeHuman` leads existing methods in multi-view consistency, geometric quality, and identity preservation. Additional examples are in Appendix Figures 6.11(a)–6.11(d).

We further validate the impact of our dataset and robust augmentation in Figure 6.5. Our dataset improves face, clothing geometry and body shape consistency.

FIGURE 6.5: **Training with our dataset achieve better (a) face consistency (b) body shape consistency (c) clothing geometry consistency. Augmentation helps with robustness to (a) half body crop (b)occlusion (c) imperfect segmentation.**

TABLE 6.3: **Quantitative results of 3D generation on CloSe.**

	SSIM ↑	LPIPS ↓	PSNR ↑	CLIP ↑	Asset CLIP ↑
DreamGaussian [115]	0.8701	0.1510	14.0147	0.7336	0.7131
LGM [116]	0.8953	0.1059	15.3836	0.8264	0.8243
SIFU [239]	0.9163	0.0897	16.6345	0.8169	0.7722
CharacterGen [33]	**0.9243**	0.0625	18.7591	0.8495	0.8566
`ReposeHuman` w/o 2DGS	0.9182	0.0793	**19.1016**	0.8590	0.8553
`ReposeHuman` w/o normal loss	0.9178	0.0632	19.0208	0.8661	0.8584
`ReposeHuman`	0.9185	**0.0613**	19.0221	**0.8672**	**0.8591**

Robust augmentation enhances model robustness to half-body crops, occlusions, and imperfect segmentation.

Takeaway. Multi-view consistency is driven by the curated disentangled dataset and dense pose controls: removing the dataset degrades all metrics, while removing dense controls leads to inconsistent side and back views. Robust augmentation has modest average gains but significantly improves performance on challenging inputs such as half-body crops and imperfect segmentation.

6.4.2 3D Disentangled Human Generation

We evaluate `ReposeHuman` against animatable human generation (CharacterGen [33]) and static human reconstruction methods (SIFU [239], DreamGaussian [115], LGM [116]). All methods use our generated canonical pose images as input.

Table 6.3 and Figure 6.6 displays the quantitative and qualitative comparisons of 3D generation in canonical poses. While CharacterGen achieves a higher SSIM, likely due to its stronger structural consistency, it fails under difficult poses, as

FIGURE 6.6: **Qualitative comparisons of 3D generation methods in original pose (left) and canonical pose (right). SIFU and LGM took our generated canonical image for canonical generation.**

shown in Figure 6.4, and struggles to maintain identity consistency. In all other metrics, our method outperforms other methods, achieving superior perceptual fidelity, color consistency, and semantic alignment. Figure 6.6 shows `ReposeHuman` has higher fidelity and better geometry for body parts such as face and limbs due to the representation using SMPL-X Gaussians. Unlike SIFU, DreamGaussian, and LGM, our models support animation. Moreover, compared to all existing models, `ReposeHuman` uniquely achieves disentanglement. 3D generation and disentanglement of assets on in the wild images are shown in Fig. 6.9. Asset transfer between different subjects are shown in Fig. 6.8.

The first row of Figure 6.7 shows that all `ReposeHuman` variants produce visually similar generation, aligning with the comparable quantitative metrics for `ReposeHuman` w/o 2DGS and `ReposeHuman` w/o normal loss in Table 6.3. In contrast, the second row reveals pronounced differences in the extracted 3D assets. Omitting 2DGS leads to poor surface coverage and degraded geometry, while removing the normal consistency loss results in noisy surface orientation, as illustrated in the last row.

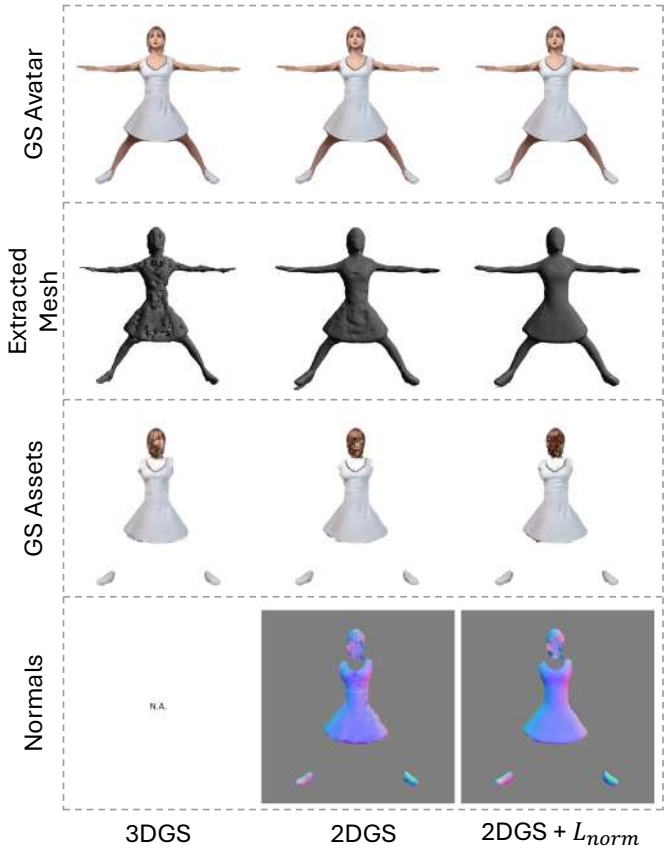


FIGURE 6.7: **Qualitative visualization of clothing geometry across representations.**

Incorporating both components yields smoother, more stable geometry, enhancing asset extraction and mesh quality. Although 2DGS and normal loss contribute modestly to image-space metrics, they are critical for high-fidelity disentangled 3D reconstruction.

Takeaway. 2D Gaussian Splatting and the normal consistency loss have little impact on image metrics but are crucial for clean, animation-ready meshes, highlighting that image metrics poorly reflect asset quality.

6.4.3 4D Human Animation and Editing

For this task, we compare our approach with other video-to-4D general GS approaches, such as DreamGaussian4D [173], as well as human-centric GS methods, including MonoHuman [177], GART [180], and GaussianAvatar [179] on the 4D-Dress dataset [236] following [228]. Figure 6.10 presents examples of animated

FIGURE 6.8: **Transfer of assets from targeting clothing to source person.**

TABLE 6.4: **CLIP-embedding loss for generated humans and segmented assets, and performance (PSNR, SSIM, LPIPS) comparison on 4D-Dress across various video-to-4D methods.**

	All ↑	Assets ↑	PSNR ↑	SSIM ↑	LPIPS ↓
DreamGaussian4D	0.784	0.769	20.54	0.93	0.080
MonoHuman	0.762	0.743	20.22	0.92	0.086
GART	0.800	0.772	18.81	0.92	0.086
GaussianAvatar	0.822	0.768	20.01	0.93	0.069
Disco4D	0.900	0.865	**25.46**	**0.96**	0.035
`ReposeHuman`	**0.908**	**0.889**	25.32	**0.96**	**0.031**

avatars generated from in-the-wild images. Table 6.4 shows that our method outperforms all others across all metrics except PSNR for Disco4D. Figure 6.7 shows that while 3DGS and 2DGS appear visually similar, our method achieves more accurate geometry than all other 3DGS-based approaches. Despite slightly outperforming Disco4D in most metrics, our approach ensures geometric accuracy using 2DGS and normal loss.

We provide extended visualizations of multi-view canonical generation in X-pose

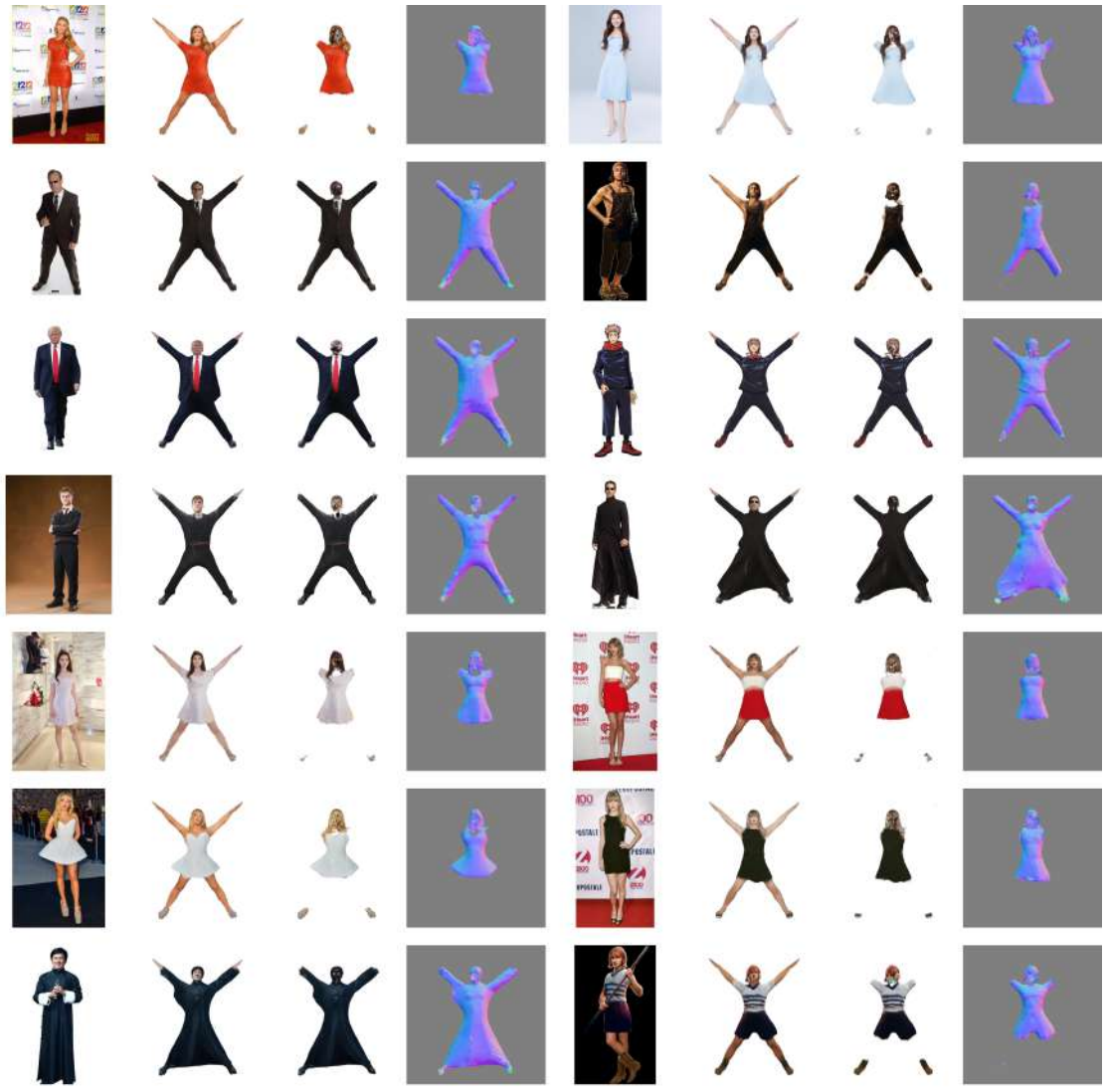


FIGURE 6.9: **Canonical 3D generation and disentanglement on in-the-wild images.**

(Fig. 6.11(d)), DA-pose (Fig. 6.11(c)), T-pose (Fig. 6.11(a)), and A-pose (Fig. 6.11(b)). Our method demonstrates robustness across different body shapes and effectively disentangles the subject from external objects, preserving structural consistency while ensuring accurate pose alignment. Additional animation and editing examples are included in Figure 6.10 .

FIGURE 6.10: **Animation and editing of our 3D avatars generated from in-the-wild images.**

6.5 Conclusion

In this work, we present `ReposeHuman`, a robust approach to generate animatable and disentangled 3D human from a single image. Unlike prior methods such as EN3D [34] and CharacterGen [33], we leverage a video diffusion model to generate

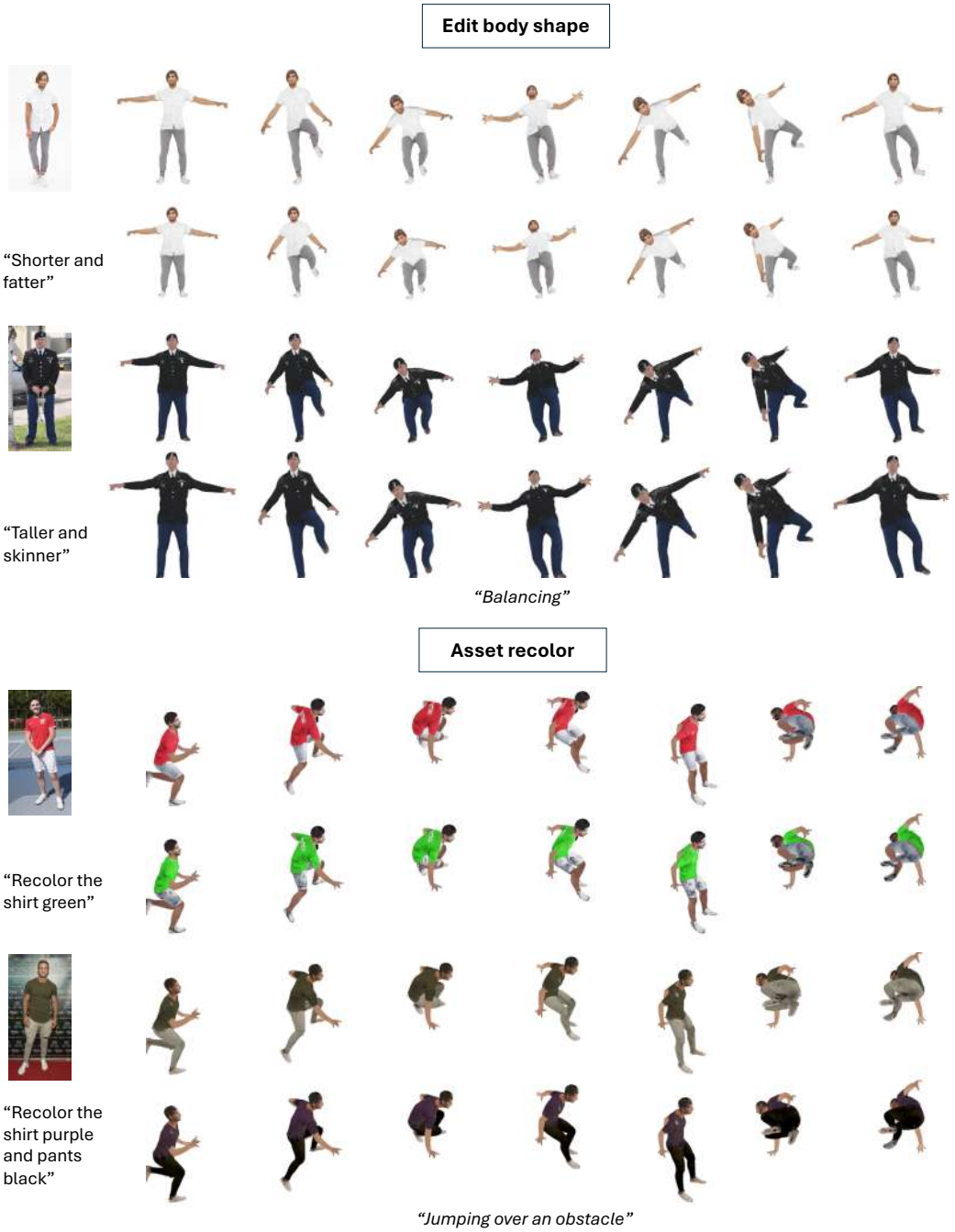
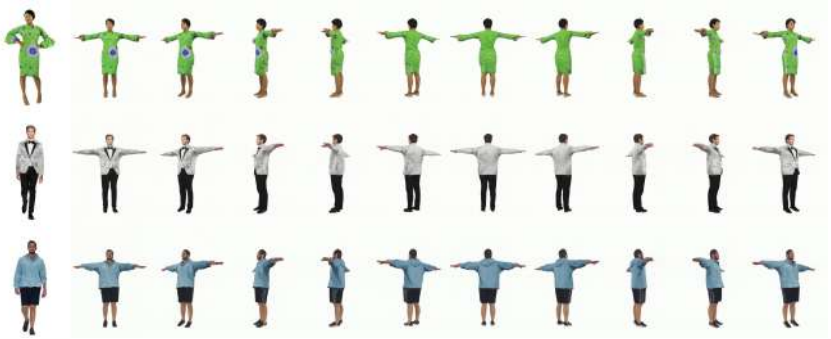
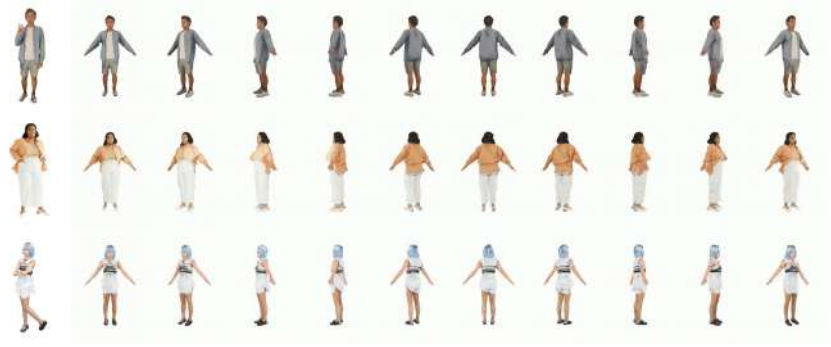


FIGURE 6.10: **Animation and editing of our 3D avatars generated from in-the-wild images.**

dense multi-view images of individuals in various canonical poses. Our approach is robust to varying body shapes and achieves superior texture and facial consistency. It also disentangles human and asset generation, enhancing flexibility for editing and composition. We further introduce a large-scale, riggable, and disentangled human dataset with diverse textures, body shapes, and clothing geometries, overcoming the



(a) **T-pose canonical outputs.**



(b) **A-pose canonical outputs.**

(c) **DA-pose canonical outputs.**

(d) **X-pose canonical outputs.**

FIGURE 6.11: **Qualitative comparison of outputs from our reposing model on SHHQ across four canonical poses: T-pose, A-pose, DA-pose, and X-pose. Each subfigure presents multi-view high-resolution outputs aligned to the respective canonical frame.**

limitations of existing datasets in clothing and texture diversity. It will be publicly available for human-centric 3D modeling research.

Limitations and Future Work. Optimizing an animatable and disentangled avatar still requires several minutes. Future work will explore feed-forward 3D reconstruction techniques to improve efficiency. Additionally, the facial resolution in the generated avatars is limited due to the relatively small pixel area of the face in the input images. To address this limitation, future work could focus on generating multi-view, high-resolution facial images to improve head reconstruction quality. Furthermore, our dataset currently lacks shoe assets, primarily due to the limited availability of open-source models. Future efforts will aim to acquire a more diverse set of shoe assets to enhance the completeness of the dataset.

Within the broader context of this thesis (Fig. 1.1), `ReposeHuman` completes the final stage by enabling canonical multi-view supervision and controllable asset generation on top of the disentangled representation from Chapter 5. Together, Chapters 3–6 form an end-to-end pipeline that converts a single in-the-wild image into an animatable 3D avatar with disentangled assets. Chapter 7 revisits this unified framework and outlines remaining challenges for future work.

Chapter 7

Conclusion and Future Works

This thesis advances the goal of achieving robust and expressive 3D human modeling by addressing key challenges across human pose and shape estimation, and generative avatar creation. Through comprehensive benchmarking, we establish strong baselines and practical insights for fair and reproducible human mesh recovery. We then propose frameworks that enhance the robustness and accuracy of whole-body pose and shape estimation, ensuring reliable predictions under challenging conditions. Moving beyond modelling, we introduce generative methods that enable compositional and disentangled human asset creation, along with large-scale datasets to support diverse and realistic human generation. Collectively, these works provide complementary contributions toward building scalable, controllable, and animatable human models, laying the foundation for future applications in animation, virtual reality, and digital content creation.

7.1 Conclusion

In the pursuit of improving 3D human mesh recovery, significant efforts have been dedicated to developing novel algorithms. However, there is a notable lack of systematic exploration into other critical factors that can impact the performance of such models. In our research, we present the *first* extensive benchmarking study, detailed in Chapter 3, that investigates various configurations for mesh recovery tasks. We thoroughly analyze these factors and identify key strategies and insights that can significantly enhance model performance. Our benchmarking study aims

to establish strong baselines that can facilitate unbiased comparisons in future mesh recovery research. A summary of our findings is provided in Section 3.7.

Moving forward to Chapter 4, we propose a novel framework called `RoboSMPLX` to advance the field of whole-body pose and shape estimation. Our framework addresses the challenges associated with accurate part localization and robustness to suboptimal part crops, resulting in reliable outputs. It introduces three innovative components: (1) an accurate subject localization module that explicitly learns sparse and dense predictions of the subject, enhancing the precision of part crop localization, (2) a robust feature extraction module that leverages supervised contrastive learning to ensure the extraction of meaningful and invariant features from suboptimal part crops, and (3) a pixel alignment module that utilizes differentiable rendering to achieve precise alignment of output pixels, leading to improved estimation of pose, shape, and camera parameters.

Chapter 5 presents Disco4D, a compositional approach for generating animatable clothed human Gaussians from a single image. By fixing SMPL-X Gaussians, fitting clothing Gaussians with identity labels over them, and leveraging diffusion models, Disco4D enables the high-fidelity reconstruction of clothed humans, including occluded parts not visible in the input image. Its disentangled representation of clothing, hair, and accessories supports flexible editing and animatability, making it highly applicable to animation and virtual asset creation.

Finally, in Chapter 6, we propose ReposeHuman, a robust pipeline that generates animatable and disentangled 3D avatars from a single image. Our key insight is to leverage a video diffusion model to synthesize multi-view canonical poses, achieving superior texture consistency and generalization to diverse body shapes. Additionally, we construct a large-scale, riggable human dataset with diverse textures, body shapes, and clothing geometries, overcoming the limitations of existing datasets in clothing and texture diversity. This dataset will be made publicly available to support future research in human-centric 3D modeling.

7.2 Future Work

Viewing this thesis as an integrated pipeline (Fig. 1.1) highlights several remaining gaps. Clothing deformation in Chapter 5 is modeled geometrically rather than

physically, limiting realism under motion. Disentangled avatar generation in Chapter 6 is further constrained by the scale and diversity of supervision, as well as by per-avatar optimisation times measured in minutes. The directions below outline natural next steps to address these limitations.

Part A: Extending the Thesis Pipeline

Direction 1: Physics-based clothing dynamics. Chapter 5 represents clothing as Gaussian layers that deform with the body via learned offsets. As a result, loose garments may appear visually broken under large or complex motions, as noted in Chapter 5's limitations. Chapter 6 extracts explicit clothing meshes from these Gaussians, but the meshes are still driven by body-conditioned deformation. Neither representation accounts for physical effects such as inertia, collision, or material properties. A natural next step is to equip both representations, the Gaussians from Chapter 5 and the meshes from Chapter 6, with a *learned* physics layer, rather than a hand-crafted physics solver. Two complementary directions are promising: (i) a lightweight differentiable physics layer that enforces collision, gravity, and material stiffness at inference time on either representation, and (ii) a learned dynamics model trained to imitate a cloth simulator on synthetic data and then applied to real-world avatars, avoiding the per-garment parameter tuning that classical simulators require. Either approach would bridge the gap between strong appearance modelling and currently limited motion realism, enabling more faithful and interactive avatar animation.

Direction 2: Scaling disentangled generation. The disentangled generation framework of Chapter 6 remains limited along three axes: efficiency, fidelity, and data scale. First, distilling the current per-avatar optimisation into a feed-forward reconstruction network could reduce inference time from minutes to seconds, enabling interactive applications. Second, introducing a face-specific super-resolution stage, conditioned on identity embeddings, could improve facial detail without increasing overall computational cost. Third, scaling the disentangled dataset using pseudo-labeling from in-the-wild video, leveraging the benchmarking methodology of Chapter 3 and the pixel-aligned estimator of Chapter 4, would significantly expand diversity in identities, clothing, and interactions. Together, these steps would transform the current framework into a scalable and deployable system.

Part B: Broader Research Trajectory

While the directions above extend the thesis pipeline, the following directions explore how the capabilities developed in this work can be applied more broadly.

Egocentric HMR for robotics. Most existing human mesh recovery methods, including those in Chapters 3 and 4, assume third-person viewpoints. In contrast, robotics applications increasingly rely on egocentric recordings, where cameras are mounted on the human body. These settings introduce new challenges such as severe truncation, motion blur, and near-field distortions. Extending robust, pixel-aligned whole-body estimation to egocentric scenarios would enable large-scale capture of human interactions for imitation learning and human-robot collaboration.

Sim-to-real human data generation. The generative models developed in Chapters 5 and 6 can also be repurposed as controllable data generators. Disentangled 3D human assets enable scalable synthesis of photorealistic data across diverse identities, clothing, and motions, particularly for scenarios that are difficult or costly to capture in the real world. A key challenge is bridging the sim-to-real gap at the video level. Combining multi-view diffusion generation with learned domain adaptation techniques offers a promising direction for producing realistic and diverse training data.

Responsible deployment. The ability to generate high-fidelity, animatable avatars from a single image introduces risks such as identity misuse, deepfake generation, and intellectual-property concerns. Addressing these challenges is essential for real-world deployment. Potential safeguards include provenance watermarking, consent-aware data pipelines, and detection models tailored to diffusion-based human generation.

Collectively, this thesis contributes strong baselines and frameworks to improve robustness in human pose and shape estimation. For human generation, it introduces compositional representations for body and clothing, as well as rich datasets that advance the field of 3D human generation and 4D animation. These contributions lay the groundwork for future research on scalable avatar creation from text or

images, offering greater personalization and fine-grained control over assets, with applications in animation, virtual reality, digital content generation, and beyond.

List of Author's Awards, Patents, and Publications

Awards

- **AISG Ph.D. Fellowship**, 2023.

Conference Proceedings

- **Hui En Pang**, Zhongang Cai, Lei Yang, Tianwei Zhang, Ziwei Liu. Benchmarking and Analyzing 3D Human Pose and Shape Estimation Beyond Algorithms. In *Proceedings of Neural Information Processing Systems (NeurIPS Dataset and Benchmark Track), 2022.*

- **Hui En Pang**, Zhongang Cai, Lei Yang, Tianwei Zhang, Qingyi Tao, Zhonghua Wu, Ziwei Liu. Towards Robust and Expressive Whole-body Human Pose and Shape Estimation. In *Proceedings of Neural Information Processing Systems, 2023.*

- Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, **Hui En Pang**, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, Ziwei Liu. SMPLer-X: Scaling Up Expressive Human Pose and Shape Estimation. In *Proceedings of Neural Information Processing Systems, (NeurIPS Dataset and Benchmark Track), 2023.*

- **Hui En Pang**, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, Ziwei Liu. Disco4D: Disentangled 4D Human Generation and Animation from a Single Image. In *Conference on Computer Vision and Pattern Recognition, 2025.*

Bibliography

[1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015. 1, 11, 16, 45, 71, 92

[2] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00744. 1, 12, 13, 15, 16, 22, 23, 24, 30, 32, 44, 45, 51, 56, 57

[3] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2252–2261, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00234. 12, 13, 14, 22, 23, 24, 32, 34, 36, 37, 44, 56, 57

[4] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11127–11137, October 2021. 12, 13, 14, 15, 16, 22, 23, 24, 28, 32, 33, 34, 36, 37, 44, 45, 51, 56, 57

[5] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, DImitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2019-June, pages 10967–10977, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.01123. 1, 11, 13, 14, 16, 23, 34, 55, 70, 71, 73, 92, 96

[6] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics from Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5614–5623, 2019. URL `https://akanazawa.github.` 1, 25, 44

[7] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. *Proceedings of the IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, pages 5252–5262, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00530. 13, 23, 24, 32

[8] Zhengyi Luo, S. Alireza Golestaneh, and Kris M. Kitani. 3D Human Motion Estimation via Motion Compression and Refinement. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12626 LNCS:324–340, 2021. ISSN 16113349. doi: 10.1007/978-3-030-69541-5_20. 1, 13, 23, 24

[9] Rıza Alp Güler and Kokkinos Iasonas. HoloPose: Holistic 3D Human Reconstruction In-The-Wild Task-Specific Decoders. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10884–10894, 2019. URL http://arielai.com/holopose. 1, 12

[10] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:4496–4505, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00463. 12, 14, 15, 36, 37

[11] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5578–5587, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00562. 12

[12] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:769–787, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_45. 12, 14, 15, 16, 45, 56

[13] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022. 12, 14, 15, 47, 52, 56, 57, 58, 60, 61, 62, 63

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, pages 42–52, 2021. doi: 10.1109/3DV53792.2021.00015. 12, 13, 14, 15, 16, 22, 24, 25, 26, 28, 31, 32, 34, 35, 37, 45, 56

[15] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic Modeling for Human Mesh Recovery. In *International Conference on Computer Vision (ICCV)*, pages 11585–11594, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01140. 12, 23, 56

[16] Sai Kumar Dwivedi, Nikos Athanasiou, Muhammed Kocabas, and Michael J. Black. Learning to Regress Bodies from Images using Differentiable Semantic Rendering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11230–11239, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01106. 12, 15, 16, 22, 23, 24, 28

[17] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11035–11045, October 2021. 1, 12, 13, 14, 15, 16, 22, 23, 24, 45

[18] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic Monocular 3D Reconstruction of Humans Wearing Clothing. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:1496–1505, 2022. ISSN 10636919. doi: 10.1109/CVPR52688.2022.00156. 1, 17, 69, 70, 90

[19] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 17, 89

[20] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 17, 89

[21] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. *Proceedings of the IEEE International Conference on Computer Vision*, pages 11026–11036, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01086. 17

[22] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3090–3099, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00316. 17

[23] Zheng Zerong, Yu Tao, Liu Yebin, and Dai Qionghai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction, 2021. 17

[24] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023. 1, 16, 17, 69, 89

[25] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, June 2022. 17, 70

[26] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 17

[27] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. In *IEEE Conference on Computer Vision (ICCV)*, 2023. 1, 17, 69, 90

[28] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4, 18, 81, 90, 100, 101

[29] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 81

[30] Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation. *arXiv*, 2024. 18, 90, 100, 101

[31] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning (ICML)*, 2025. 4, 100

[32] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance, 2024. 4, 18, 81, 90, 96, 100, 101

[33] Hao-Yang Peng, Jia-Peng Zhang, Meng-Hao Guo, Yan-Pei Cao, and Shi-Min Hu. Charactergen: Efficient 3d character generation from single images with multi-view pose canonicalization. *ACM Transactions on Graphics (TOG)*, 43 (4), 2024. doi: 10.1145/3658217. 4, 18, 90, 96, 100, 101, 102, 107

[34] Yifang Men, Biwen Lei, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. En3d: An enhanced generative model for sculpting 3d humans from 2d synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4, 18, 90, 96, 107

[35] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 8, 91, 93, 95

[36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 8, 91, 93, 96

[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation, 2016. URL https://arxiv.org/abs/1603.06937. 11

[38] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *European Conference on Computer Vision (ECCV)*, pages 598–613, 2020. URL https://star.is.tue.mpg.de. 11

[39] Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Neural descent for visual 3D human pose and shape. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 14479–14488, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01425. 11, 12, 13, 34, 37

[40] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813. 11, 45

[41] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017. 12, 45

[42] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3D human mesh from monocular images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(12):15406–15425, 2023. 12, 24, 26, 58

[43] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6M. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2014. ISSN 01628828. URL http://109.101.234.42/documente/publications/1-82.pdf. 12, 13, 23, 24, 25, 26, 35, 37, 55

[44] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, pages 506–516, 2017. doi: 10.1109/3DV.2017.00064. 12, 23, 25, 27, 37

[45] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in*

Bioinformatics), 8693 LNCS(PART 5):740–755, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10602-1_48. 12, 25, 55

[46] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1465–1472, 2011. ISSN 10636919. doi: 10.1109/CVPR.2011.5995318. 12, 25

[47] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12. 12, 25

[48] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. ISSN 10636919. doi: 10.1109/CVPR.2014.471. 12, 25, 55

[49] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. *arXiv preprint arXiv:2204.13686*, 2022. 12

[50] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pages 120–130, 2018. doi: 10.1109/3DV.2018.00024. 12, 13, 23, 25, 27, 34, 37

[51] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00542. 12, 23, 25

[52] Song Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi Min Hu. Pose2Seg: Detection free human instance segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:889–898, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00098. 12, 23, 25

[53] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. *Proceedings - 2018 International Conference on 3D Vision, 3DV 2018*, pages 484–494, 2018. doi: 10.1109/3DV.2018.00062. 12, 15

[54] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-Occluded Human Shape and Pose Estimation from a Single Color Image. *Proceedings of the IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, pages 7374–7383, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00740. 12, 15, 25

[55] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical Kinematic Human Mesh Recovery. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12362 LNCS: 768–784, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58520-4_45. 12, 13, 34

[56] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel Prediction Network for Accurate 3D Human Pose and Mesh Estimation from a Single RGB Image. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12352 LNCS:752–768, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58571-6_44. 12, 56, 58

[57] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 12, 15, 16, 45, 56

[58] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3382–3392, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00339. 12, 22, 24, 34, 37

[59] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-End Human Pose and Mesh Reconstruction with Transformers. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00199. 12, 13, 14, 24, 45, 56, 58

[60] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body Meshes as Points. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 546–556, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00061. 12, 13, 34

[61] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 12, 24

[62] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh Graphormer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12919–12928, 2021. ISBN 9781665428125. doi: 10.1109/ICCV48922.2021.01270. 12, 24, 36, 37

[63] Mihai Zanfir, Andrei Zanfir, Eduard Gabriel Bazavan, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. THUNDR: Transformer-based 3D HUmaN Reconstruction with Markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022. 12

[64] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *International Conference on Computer Vision (ICCV)*, pages 11426–11436, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.01125. 12, 22, 24

[65] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11214 LNCS:614–631, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01249-6_37. 13, 23, 24, 25, 27, 35, 55

[66] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly Supervised 3D Human Pose and Shape Reconstruction with Normalizing Flows. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12351 LNCS:465–481, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58539-6_28. 13, 15, 24

[67] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic 3D human shape and pose estimation from multiple unconstrained images in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 16089–16099, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.01583.

[68] Lee Gun-Hee and Lee Seong-Whan. Uncertainty-Aware Human Mesh Recovery from Video by Learning Part-Based 3D Dynamics. *ICCV*, pages 12375–12384, 2021. 13, 24

[69] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. ISBN 9781467383912. doi: 10.1109/ICCV.2015.381. 13

[70] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 13, 23, 24, 25, 26, 55

[71] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9908 LNCS:630–645, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46493-0_38. 13, 32

[72] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:5686–5696, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.00584. 13, 32

[73] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Jiatong Li, Zhengyu Lin, Haiyu Zhao, Shuai Yi, Lei Yang, Chen Change Loy, and Ziwei Liu. Playing for 3D Human Recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2024. 13, 24, 26, 28, 32

[74] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 13

[75] Chris Rockwell and David F. Fouhey. Full-Body Awareness from Partial Observations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12362 LNCS:522–539, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58520-4_31. 13, 34

[76] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. In *British Machine Vision Conference (BMVC)*, September 2020. 34

[77] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3D human pose estimation: Motion to the rescue. *Advances in Neural Information Processing Systems*, 32, 2019. ISSN 10495258.

[78] Yuanlu Xu, Song Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:7759–7769, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00785. 13, 15, 16, 34

[79] Adnane Boukhayma, Rodrigo De Bem, and Philip H.S. Torr. 3D hand shape and pose from images in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 10835–10844, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01110. 14, 45

[80] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:285–295, 2019. ISSN 21607516. doi: 10.1109/CVPRW.2019.00038. 14, 45

[81] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total Capture: A 3D Deformation Model for Tracking Faces, Hands, and Bodies. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00868. 14

[82] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 10957–10966, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01122.

[83] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM GHUML: Generative 3D human shape and articulated pose models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6183–6192, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00622. 14

[84] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 14, 47, 60, 61, 62

[85] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://expose.is.tue.mpg.de. 14, 16, 45, 47, 50, 56, 57, 58, 59, 60, 61, 62

[86] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2021-Octob, pages 1749–1759, 2021. ISBN 9781665401913. doi: 10.1109/ICCVW54120.2021.00201. 14, 56

[87] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *International Conference on 3D Vision (3DV)*, 2021. 14, 16, 45, 47, 50, 55, 56, 60, 61, 62

[88] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-Supervised 3D Hand Pose Estimation from monocular RGB via Contrastive Learning. *Proceedings of the IEEE International Conference on Computer Vision*, pages 11210–11219, 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01104. 15, 62, 64

[89] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive Representation Learning for Hand Shape Estimation. *Lecture Notes in Computer*

Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13024 LNCS:250–264, 2021. ISSN 16113349. doi: 10.1007/978-3-030-92659-5_16. 15, 45, 64

[90] Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Rethinking Self-Supervised Visual Representation Learning in Pre-training for 3D Human Pose and Shape Estimation. *International Conference on Learning Representations (ICLR)*, pages 1–18, 2023. URL http://arxiv.org/abs/2303.05370. 15, 54, 62, 63

[91] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 15, 54

[92] Hsiao Yu Fish Tung, Hsiao Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. *Advances in Neural Information Processing Systems*, 2017-December(Nips):5237–5247, 2017. ISSN 10495258. 15

[93] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems*, 2024. 15

[94] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts, 2022. 16

[95] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions, 2023. 16

[96] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 16

[97] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 16

[98] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[99] Shivam Duggal and Deepak Pathak. Topologically-aware deformation fields for single-view 3d reconstruction. *CVPR*, 2022.

[100] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction, 2023.

[101] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 16

[102] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 16

[103] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8446–8455, 2023.

[104] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22819–22829, October 2023.

[105] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 16, 18, 90

[106] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 16, 19, 77, 99

[107] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[108] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 17

[109] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=0jHkUDyE09. 16

[110] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 16

[111] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 16, 18, 90

[112] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024. 16

[113] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 16, 19

[114] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/. 16, 92

[115] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *International Conference on Learning Representations (ICLR)*, 2024. 16, 17, 72, 79, 85, 87, 102

[116] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view Gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision (ECCV)*, 2024. 16, 17, 72, 73, 79, 85, 102

[117] Xuanyi Li, Daquan Zhou, Chenxu Zhang, Shaodong Wei, Qibin Hou, and Ming-Ming Cheng. Sora generates videos with stunning geometrical consistency. *arXiv preprint arXiv: 2402.17403*, 2024. 16

[118] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 73

[119] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models, 2023.

[120] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman,

Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

[121] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

[122] Yan Zeng, Guoqiang Wei, Jiani Zheng, Jiaxin Zou, Yang Wei, Yuchen Zhang, and Hang Li. Make pixels dance: High-dynamic video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[123] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation, 2024. 16

[124] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators, 2024. 16

[125] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4491–4500, 2019. 16

[126] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *Proceedings of International Conference on 3D Vision (3DV '20)*, pages 322 – 332, November 2020. 16

[127] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *European Conference on Computer Vision*. Springer, 2020. 16, 17

[128] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 16

[129] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images, 2019. 16

[130] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14718–14727, October 2021. 16

[131] Yuxuan Xue, Xianghui Xie, Riccardo Marin, and Gerard. Pons-Moll. Human 3Diffusion: Realistic Avatar Creation via Explicit 3D Consistent Diffusion Models. In *Advances in Neural Information Processing Systems*, 2024. 17

[132] Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, and Zhaoxin Fan. D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2023. 17

[133] Yifan Yang, Dong Liu, Shuhai Zhang, Zeshuai Deng, Zixiong Huang, and Mingkui Tan. Hilo: Detailed and robust 3d clothed human reconstruction with high-and low-frequency information of parametric models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10671–10681, 2024. 17

[134] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *European Conference on Computer Vision*, pages 344–359. Springer, 2020. 17

[135] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.

[136] Igor Santesteban, Miguel A. Otaduy, and Dan Casas. Learning-Based Animation of Clothing for Virtual Try-On. *Computer Graphics Forum (Proc. Eurographics)*, 2019. ISSN 1467-8659. doi: 10.1111/cgf.13643.

[137] Raquel Vidaurre, Igor Santesteban, Elena Garces, and Dan Casas. Fully Convolutional Graph Neural Networks for Parametric Virtual Try-On. *Computer Graphics Forum (Proc. SCA)*, 2020. 17

[138] Xin Chen, Anqi Pang, Yang Wei, Wang Peihao, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (Presented at ACM SIGGRAPH)*, 2021. 17

[139] Oshri Halimi, Fabian Prada, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, and Yaser Sheikh. Garment avatars: Realistic cloth driving using pattern registration, 2022.

[140] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. URL http://dx.doi.org/10.1145/3072959.3073711. Two first authors contributed equally.

[141] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics*, 40 (6):1–15, December 2021. ISSN 1557-7368. doi: 10.1145/3478513.3480545. URL http://dx.doi.org/10.1145/3478513.3480545. 17

[142] Zhu Heming, Cao Yu, Jin Hang, Chen Weikai, Du Dong, Wang Zhangye, Cui Shuguang, and Han Xiaoguang. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision – ECCV 2020*, pages 512–530. Springer International Publishing, 2020. ISBN 978-3-030-58452-8. 17

[143] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 18

[144] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J. Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, 2022. 18, 70, 90

[145] Yao Feng, Weiyang Liu, Timo Bolkart, Jinlong Yang, Marc Pollefeys, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. *arXiv*, 2023. 18, 70, 90

[146] Xu He, Xiaoyu Li, Di Kang, Jiangnan Ye, Chaopeng Zhang, Liyang Chen, Xiangjun Gao, Han Zhang, Zhiyong Wu, and Haolin Zhuang. Magicman: Generative novel view synthesis of humans with 3d-aware diffusion and iterative refinement, 2024. 18, 90

[147] Zhibin Liu, Haoye Dong, Aviral Chharia, and Hefeng Wu. Human-vdm: Learning single-image 3d human gaussian splatting from video diffusion models, 2024. URL https://arxiv.org/abs/2409.02851. 18, 90

[148] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *CVPR*, 2023. 18

[149] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 18

[150] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, June 2021. 18, 96

[151] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2023)*, June 2023. 18

[152] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 19

[153] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 19

[154] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[155] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 19

[156] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 19

[157] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[158] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[159] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021.

[160] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[161] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2021.

[162] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13144–13152, 2021. 19

[163] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 19

[164] Sara Fridovich-Keil and Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023.

[165] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 19

[166] Haithem Turki, Jason Y. Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes, 2023. 19

[167] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22. ACM, November 2022. doi: 10.1145/3550469.3555383. URL http://dx.doi.org/10.1145/3550469.3555383. 19

[168] Jad Abou-Chakra, Feras Dayoub, and Niko Sünderhauf. Particlenerf: Particle based encoding for online neural radiance fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[169] Shanyan Guan, Huayu Deng, Yunbo Wang, and Xiaokang Yang. Neurofluid: Fluid dynamics grounding with particle-driven neural radiance fields, 2022. 19

[170] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 19

[171] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 19

[172] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 19

[173] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 19, 77, 82, 83, 99, 104

[174] Muhammed Kocabas, Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats, 2023. URL https://arxiv.org/abs/2311.17910. 19

[175] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024.

[176] Yang Liu, Xiang Huang, Minghan Qin, Qinwei Lin, and Haoqian Wang. Animatable 3d gaussian: Fast and high-quality reconstruction of multiple human avatars. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, pages 1120–1129, 2024.

[177] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 82, 104

[178] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. 2024.

[179] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 82, 104

[180] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models, 2023. URL https://arxiv.org/abs/2311.16099. 19, 82, 104

[181] Hui En Pang, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Benchmarking and analyzing 3d human pose and shape estimation beyond algorithms. In *NeurIPS*, 2022. 21, 26, 44, 59

[182] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking People by Predicting 3D Appearance, Location & Pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. URL http://arxiv.org/abs/2112.04477. 23

[183] Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, and Jitendra Malik. Tracking people with 3d representations. In *NeurIPS*, 2021. 23

[184] Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Human Mesh Recovery from Multiple Shots. In *Computer Vision and Pattern Recognition*, 2022. URL http://arxiv.org/abs/2012.09843. 23

[185] Ren Li, Meng Zheng, Srikrishna Karanam, Terrence Chen, and Ziyan Wu. Everybody Is Unique: Towards Unbiased Human Mesh Recovery. In *British Machine Vision Conference*, pages 1–13, 2021. URL http://arxiv.org/abs/2107.06239. 24

[186] Mohamed Hassan, Vasileios Choutas, DImitrios Tzionas, and Michael Black. Resolving 3D human pose ambiguities with 3D scene constraints. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2282–2292, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00237. 25, 27

[187] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into Person: Self-supervised Structure-sensitive Learning and a new benchmark for human parsing. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6757–6765, 2017. doi: 10.1109/CVPR.2017.715. 25

[188] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:10855–10864, 2019. ISSN 10636919. doi: 10.1109/CVPR.2019.01112. 25

[189] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Weng, and Yonggang Wang. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. In *IEEE International Conference on Multimedia and Expo*, pages 1480–1485, 2017. ISBN 9781538695524. doi: 10.1109/ICME.2019.00256. 25

[190] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *IEEE International Conference on Computer Vision*, pages 2248–2255, 2013. ISBN 9781479928392. doi: 10.1109/ICCV.2013.280. 25

[191] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-Body Human Pose Estimation in the Wild. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12354 LNCS: 196–214, 2020. ISSN 16113349. doi: 10.1007/978-3-030-58545-7_12. 25, 55

[192] Albert Pumarola, Jordi Sanchez, Gary P.T. Choi, Alberto Sanfeliu, and Francesc Moreno. 3Dpeople: Modeling the geometry of dressed humans. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob:2242–2251, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00233. 25

[193] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4627–4635, 2017. doi: 10.1109/CVPR.2017.492. 25

[194] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and

2D human representations. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:4704–4713, 2017. doi: 10.1109/CVPR.2017.500. 25

[195] Lea Müller, Ahmed A.A. Osman, Siyu Tang, Chun Hao P. Huang, and Michael J. Black. On Self-Contact and Human Pose. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9985–9994, 2021. ISSN 10636919. doi: 10.1109/CVPR46437.2021.00986. 25

[196] Michał Rapczyński, Philipp Werner, Sebastian Handrich, and Ayoub Al-Hamadi. A baseline for cross-database 3d human pose estimation. *Sensors*, 21(11), 2021. ISSN 14248220. doi: 10.3390/s21113769. 26

[197] Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, and Chen Change Loy. Messytable: Instance association in multiple camera views. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 26

[198] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 26

[199] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *35th International Conference on Machine Learning, ICML 2018*, 10:6900–6909, 2018. 30

[200] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in Neural Information Processing Systems*, 32(NeurIPS): 1–12, 2019. ISSN 10495258. 30

[201] Daniel Holden. Robust solving of optical motion capture data by denoising. *ACM Transactions on Graphics*, 37(4):1–12, 2018. ISSN 15577368. doi: 10.1145/3197517.3201302. 31

[202] Joela F. Gauss, Christoph Brandin, Andreas Heberle, and Welf Löwe. Smoothing Skeleton Avatar Visualizations Using Signal Processing Technology. *SN Computer Science*, 2(6):1–17, 2021. ISSN 26618907. doi: 10.1007/s42979-021-00814-2. URL https://doi.org/10.1007/s42979-021-00814-2. 31

[203] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, 2019. 32

[204] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 32

[205] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. 32

[206] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. ISBN 9781665428125. doi: 10.1109/iccv48922.2021.00986. 32

[207] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. In *NeurIPS*, pages 1–14, 2021. URL http://arxiv.org/abs/2104.13840. 32

[208] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 13001–13008, 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i07.7000. 34

[209] Minghui Chen, Zhiqiang Wang, and Feng Zheng. Benchmarks for corruption invariant person re-identification, 2021. 34

[210] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020. 34

[211] Zhang Feng, Xiatian Zhu, and Mao Ye. Fast Pose Estimation. In *Computer Vision and Pattern Recognition*, pages 3517–3526, 2019. 35

[212] Tewodros Legesse Munea, Yalew Zelalem Jembre, Halefom Tekle Weldegebriel, Longbiao Chen, Chenxi Huang, and Chenhui Yang. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access*, 8:133330–133348, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3010248. 35

[213] Aritra Ghosh, Naresh Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015. ISSN 18728286. doi: 10.1016/j.neucom.2014.09.081. 35

[214] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *European Conference on Computer Vision*, 2022. URL http://arxiv.org/abs/2208.00571. 37

[215] Hui En Pang, Zhongang Cai, Lei Yang, Tao Qingyi, Wu Zhonghua, Tianwei Zhang, and Ziwei Liu. Towards robust and expressive whole-body human pose and shape estimation. In *NeurIPS*, 2023. 45, 47, 58, 59

[216] Radek Danecek, Michael Black, and Timo Bolkart. EMOCA: Emotion Driven Monocular Face Capture and Animation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June: 20279–20290, 2022. ISSN 10636919. doi: 10.1109/CVPR52688.2022.01967. 45

[217] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, volume 40, 2021. URL https://doi.org/10.1145/3450626.3459936. 45

[218] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular Real-time Full Body Capture with Inter-part Correlations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, number 61822111 in 61822111, pages 4809–4820, 2021. ISBN 9781665445092. doi: 10.1109/CVPR46437.2021.00478. 45, 56

[219] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 50

[220] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 50

[221] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11210 LNCS:536–553, 2018. ISSN 16113349. doi: 10.1007/978-3-030-01231-1_33. 53

[222] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2022-June:2298–2306, 2022. ISSN 21607516. doi: 10.1109/CVPRW56347.2022.00256. 55

[223] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-October:813–822, 2019. ISSN 15505499. doi: 10.1109/ICCV.2019.00090. 55, 56

[224] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 55

[225] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228, 2021. ISSN 19393539. doi: 10.1109/TPAMI.2020.2970919. 55

[226] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 9319–9328, 2020. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00934. 55

[227] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019. ISSN 19493045. doi: 10.1109/TAFFC.2017.2740923. 55

[228] Hui En Pang, Shuai Liu, Zhongang Cai, Lei Yang, Tianwei Zhang, and Ziwei Liu. Disco4d: Disentangled 4d human generation and animation from a single image, 2024. URL https://arxiv.org/abs/2409.17280. 69, 75, 98, 100, 104

[229] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2023. 73, 98

[230] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 73

[231] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 74

[232] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:*, 2023. 74

[233] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR*, abs/2105.15203, 2021. URL https://arxiv.org/abs/2105.15203. 75

[234] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision (ECCV)*, 2024. 75, 98

[235] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023. 75

[236] Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4d-dress: A 4d dataset of real-world human clothing with semantic annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 78, 81, 96, 97, 100, 104

[237] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20282–20292, October 2023. 79, 96

[238] Dimitrije Antić, Garvita Tiwari, Batuhan Ozcomlekci, Riccardo Marin, and Gerard Pons-Moll. CloSe: A 3D clothing segmentation dataset and model. In *International Conference on 3D Vision (3DV)*, March 2024. 79, 96, 97, 100

[239] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9936–9947, June 2024. 89, 102

[240] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096*, 2023. 90

[241] Siyou Lin, Zhe Li, Zhaoqi Su, Zerong Zheng, Hongwen Zhang, and Yebin Liu. Layga: Layered gaussian avatars for animatable clothing transfer. In *SIGGRAPH Conference Papers*, 2024. 90

[242] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. doi: 10.1145/3641519.3657428. 91, 99

[243] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations (ICLR)*, 2025. 92, 94

[244] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html. 92, 94

[245] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. 92, 95

[246] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision (ECCV)*, 2024. 93, 95

[247] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, June 2023. 96

[248] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2023. 96

[249] Maria Korosteleva and Olga Sorkine-Hornung. GarmentCode: Programming parametric sewing patterns. *ACM Transaction on Graphics*, 42(6), 2023. doi: 10.1145/3618351. SIGGRAPH ASIA 2023 issue. 97

[250] Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, and Justus Thies. Text-conditioned generative model of 3d strand-based human hairstyles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4703–4712, June 2024. 97

[251] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 97