MVHumanNet: A Large-scale Dataset of Multi-view Daily Dressing Human Captures

Zhangyang Xiong[1,2#]   Chenghong Li[1,2#]   Kenkun Liu[2#]   Hongjie Liao[2]   Jianqiao Hu[2]
Junyi Zhu[2]   Shuliang Ning[1,2]   Lingteng Qiu[2]   Chongjie Wang[2]   Shijie Wang[2]
Shuguang Cui[2,1]   Xiaoguang Han[2,1*]

[#]equal contribution   [*]corresponding author
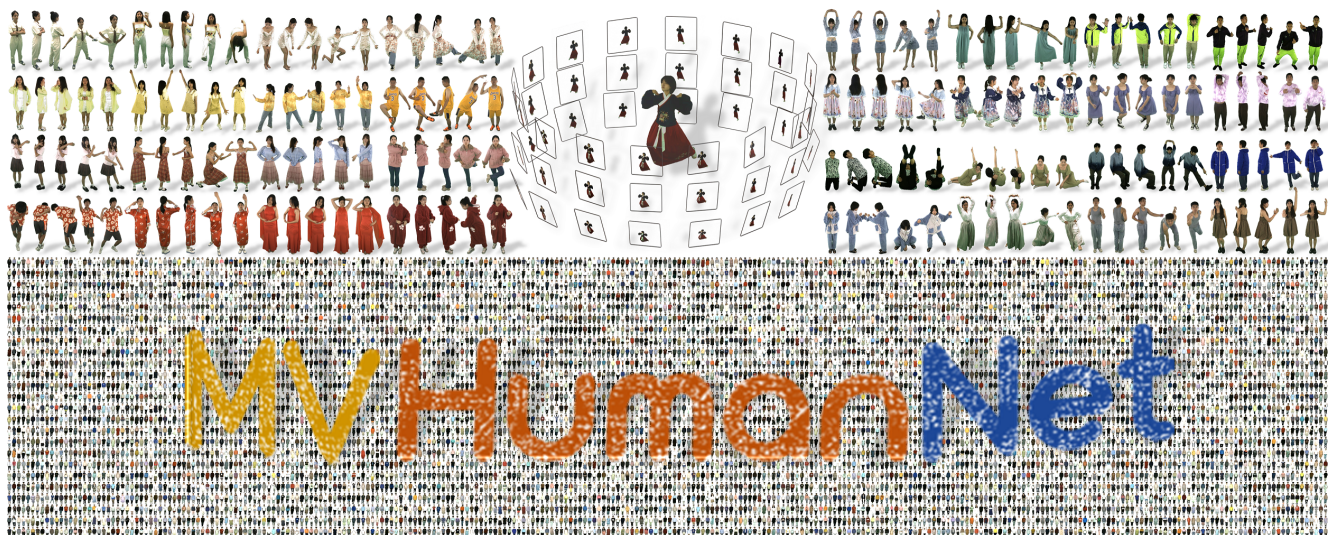[1]FNii, CUHKSZ   [2]SSE, CUHKSZ

Figure 1. We introduce **MVHumanNet**, a large-scale dataset of *multi-view human images* with unprecedented scale in human subjects, daily outfits, motion sequences and frames. **Top left and right**: Examples of multi-view poses featuring different human identities with various daily dressing in our dataset. **Top middle**: Our multi-view capture system includes 48 cameras of 12MP resolution. **Bottom**: Comprehensive visualization of all 9000 outfits in our MVHumanNet. Refer to GAP-LAB-CUHK-SZ/MVHumanNet.

Abstract

*In this era, the success of large language models and text-to-image models can be attributed to the driving force of large-scale datasets. However, in the realm of 3D vision, while remarkable progress has been made with models trained on large-scale synthetic and real-captured object data like Objaverse and MVImgNet, a similar level of progress has not been observed in the domain of human-centric tasks partially due to the lack of a large-scale human dataset. Existing datasets of high-fidelity 3D human capture continue to be mid-sized due to the significant challenges in acquiring large-scale high-quality 3D human data. To bridge this gap, we present **MVHuman-Net**, a dataset that comprises multi-view human action sequences of **4,500** human identities. The primary focus of our work is on collecting human data that features a large number of diverse identities and everyday clothing using a multi-view human capture system, which facilitates easily scalable data collection. Our dataset contains **9,000** daily outfits, **60,000** motion sequences and **645 million** frames with extensive annotations, including human masks, camera parameters, 2D and 3D keypoints, SMPL/SMPLX parameters, and corresponding textual descriptions. To explore the potential of MVHumanNet in various 2D and 3D visual tasks, we conducted pilot studies on view-consistent action recognition, human NeRF reconstruction, text-driven view-unconstrained human image generation, as well as 2D view-unconstrained human image and 3D avatar generation. Extensive experiments demonstrate the performance improvements and effective applications enabled by the scale provided by MVHumanNet. As the current largest-scale 3D human dataset, we hope that the release of MVHumanNet data with annotations will foster further innova-*

tions in the domain of 3D human-centric tasks at scale.

1. Introduction

In recent years, the exponential advancements of AI have been largely driven by the massive amounts of data. In computer vision, with the emergency of SA-1B [38] and LAION-5B [62], models like SAM [38] and Stable Diffusion [58] have greatly benefited from these large volumes of data, enabling zero-shot transfer to downstream tasks. Subsequently, Objaverse [18, 19] and MVImgNet [75] break barriers of 3D data collection with large-scale synthetic 3D assets and real-world multi-view capture, which support Zero123 [45] and LRM [29] models to achieve impressive generalization ability of multi-view or 3D reconstruction. However, comparable progress on human-centric tasks still remained elusive due to the limited scale of 3D human data.

Compared to collecting 3D object datasets, capturing high-quality and large-scale 3D human avatars is more time-consuming in the same order of scale. Existing 3D human datasets can be categorized into two distinct representations: 3D human scans and multi-view human images. While 3D human scan data [57, 74] provides accurate geometric shapes, it comes with high acquisition costs which leads to limited data scale. Conversely, multi-view capture provides an easier way to collect 3D human data. Previous multi-view human datasets [31, 41, 55] involve only a few dozen human subjects. Recent advances in multi-view human performance data [12, 13] narrow the gap of data scarcity which provides more diverse and representative human data for establishing reasonable benchmarks. To ensure comprehensiveness, it is necessary for these datasets to consider the complex clothing and the uncommon human-object interaction. However, incorporating these factors introduces complexities for scaling up the dataset.

To scale up the 3D human data, we present **MVHuman-Net**, a large-scale multi-view human performance capture dataset. Our dataset primarily focuses on casual clothing commonly found in everyday life, enabling to easily expand the scale of human data collection. For the hardware setup, we establish two 360-degree indoor systems equipped with 48 and 24 calibrated RGB cameras, respectively, to capture high-fidelity videos with resolutions up to 12MP (4096 × 3000) and 5MP (2048 × 2448). Considering the capture of human data, we intend to cover a wide range of attributes among human subjects, including age, body shape, motion, as well as the colors, types, and materials of dressing, enabling our dataset as diverse as possible. Furthermore, we also design 500 motion types to guarantee coverage of daily scenarios. Overall, we invite **4,500** individuals to participate in the data capture process. Each participant is recorded in two distinctive outfits (**9,000** in total) and at least seven different motion sequences. Thanks to the targeted collection of everyday clothing, data capture for each participant has been accomplished efficiently within six months. Eventually, the full dataset comprises an extensive collection of **60,000** motion sequences with over **645 million** frames. Compared with the existing multi-view human datasets [12, 30, 31, 55], MVHumanNet provides a significantly larger number of human subjects and outfits than previously available. Furthermore, MVHumanNet surpasses the recently proposed DNA-Rendering [13] dataset by an order of magnitude in terms of motion and frame data. The detailed comparisons between MVHumanNet and other relevant datasets are shown in Tab. 1.

In order to benefit downstream human-centric tasks, we also provide essential annotations including action labels, camera intrinsics and extrinsics, human masks, 2D/3D keypoints, SMPL/SMPLX parameters and text descriptions to enhance the applicability of our dataset. To thoroughly explore the capabilities of our dataset, we carefully design four pilot experiments: **a)** view-consistent action recognition, **b)** NeRF reconstruction for human, **c)** text-driven view-unconstrained human image generation, and **d)** 2D view-unconstrained human image and 3D avatar generation. First, leveraging the multi-view nature of human capture data, we can achieve more accurate view-consistent action recognition and enhance the generalization capability of NeRF as the data scale increases. Furthermore, the unprecedented scale of subjects and outfits, along with pose sequences and paired textual descriptions, allows us to finetune a remarkable text-driven, pose-conditioned high-quality human image generation model. Finally, through the exploitation of multi-view human images on a large scale, we can obtain 2D/3D full-body human generative models with promising results. The aforementioned experiments reveal the promise and opportunities with the large-scale MVHumanNet datasets to boost a wide range of digital human applications and inspire future research.

In summary, the main contributions of our work include:

- We present the largest multi-view human capture dataset, which is nearly ten times larger than the recently proposed DNA-Rendering dataset in terms of human subjects, motion sequences, and frames.

- We conduct several pilot studies that demonstrate the proposed MVHumanNet can support various downstream human-centric tasks for effective applications.

- We believe that MVHumanNet opens up new possibilities for research in the field of 3D digital human.

2. Related Work

3D Human Reconstruction and Generation. Recently, we have witnessed impressive performance in the field of image generation, 3D reconstruction and novel view synthesis in computer vision community with the emergency

Dataset	Age	Cloth	Motion	#ID	#Outfit	#Actions	#View	#Frames	Resolution
Human3.6M [31]	✗	✗	✔	11	11	17	4	3.6M	1000P
CMU Panoptic [35]	✔	✗	✔	97	97	65	31	15.3M	1080P
MPI-INF-3DHP [50]	✗	✗	✔	8	8	–	14	1.3M	2048P
NHR [71]	✗	✗	✔	3	3	5	80	100K	2048P
ZJU-MoCap [55]	✗	✗	✔	10	10	10	24	180K	1024P
Neural Actor [44]	✗	✗	✔	8	8	–	11∼100	250K	1285P
HUMBI [76]	✔	✔	✗	772	772	–	107	26M	1080P
AIST++ [41]	✗	✗	✗	30	30	–	9	10.1M	1080P
THuman 4.0 [81]	✗	✗	✔	3	3	–	24	10K	1150P
HuMMan [5]	✗	✔	✔	1000	1000	500	10	60M	1080P
GeneBody [12]	✔	✔	✔	50	100	61	48	2.95M	2048P
ActorsHQ [30]	✗	✗	✔	8	8	52	160	40K	4096P
DNA-Rendering [13]	✔	✔	✔	500	1500	1187	60	67.5M	4096P
MVHumanNet(Ours)	✔	✔	✔	4500	9000	500	48	645.1M	4096P

Table 1. **Dataset comparison on existing multi-view human-centric datasets.** MVHumanNet provides a significantly larger number of human subjects and outfits than previous datasets available, regarding the number of identities (#ID), outfits in total (#Outfit) and frames of images (#Frames). Attributes among humans, including age, cloth and motion are covered (denoted by ✔ for inclusion and ✗ for exclusion.). Cells highlighted in ▉ ▉ denotes the dataset with best and second-best feature in each column.

of Generative Adversarial Networks (GANs) [23, 32, 36], Neural Implicit Function [11, 51, 54] and Neural Radiance Field (NeRF) [52, 53]. These successes inspire subsequent works [20, 40, 55, 60] to extend reconstruction and generation tasks to high-fidelity clothed full-body humans. Many efforts have also been made to combine 2D GANs with NeRF representations for 3D-aware, photo-realistic image synthesis. EG3D [7] proposes the 3D-aware generation of multi-view face images by introducing an efficient tri-plane representation for volumetric rendering. GET3D [22] utilizes two separate latent codes to generate the SDF and texture field, enabling the generation of textured 3D meshes. EVA3D [27] extends EG3D to learn generative models with human body priors for 3D full-body human generation from a collection of 2D images. HumanGen [33] and Get3DHuman [72] further incorporate the priors of StyleGAN-Human [21] and PIFuHD [61] for generative human model construction. In addition, Text2Human [34] and AvatarClip [28] explore to leverage the powerful vision-language model CLIP [56] for text-driven 2D and 3D human generation. However, the reconstruction, generation and novel view synthesis works can only utilize limited real-world human data, which consequently affects the generalizability of their models. Moreover, the current methods of human generation often train their models on datasets comprising only front-view 2D human images [21, 46] or monocular human videos [77]. Unfortunately, these approaches fail to produce satisfactory results when altering the input image across various camera viewpoints. In this work, we provide the current largest scale of multi-view human capture images along with text descriptions to facilitate 3D human-centric tasks.

3D Human Scanning Datasets. Understanding human actions and reconstructing detailed body geometries with realistic appearances are challenging tasks that require high-quality and large-scale human data. Early works [3, 4, 78] in this field provide dynamic human scans but with limited data consisting of only a few subjects or simple postures. Parallel works such as Northwestern-UCLA [69] and NTU RGB+D series [43, 63] utilize more affordable Kinect sensors to obtain depth and human skeleton data, enabling the capture of both appearance and action cues. However, due to the limitations in the accuracy of Kinect sensors, these datasets are inadequate for precise human body modeling. Subsequently, AMASS [49] further integrates traditional motion capture datasets [15, 59] and expands them with fully rigged 3D meshes to facilitate advancements in human action analysis and body modeling research. With the emergency of learning-based digital human techniques, relevant algorithms [9, 60, 61, 73] heavily rely on human scan datasets with high-fidelity 3D geometry and corresponding images. Several studies [26, 48, 64, 74, 79, 80] capture their own datasets and release the data to the public for research purposes. Additionally, there are several commercial scan datasets [1, 2, 57, 66] that are well-polished and used for research to ensure professional quality. These datasets play a foundational role in bridging the gap between synthetic virtual avatars and real humans. However, the aforementioned datasets typically exhibit a bias towards standing poses due to the complicated capture procedure and cannot afford for

large-scale data collection.

Multi-view Human Capturing Datasets. Multi-view capture holds an indispensable role in computer vision, serving as a fundamental technique for AR/VR and 3D content production. Prior works [67, 68] present multi-view stereo systems to collect multi-view human images and apply multi-view constraints to reconstruct 3D virtual characters. Human3.6M [31] captures numerous 3D human poses using a marker-based motion capture system from 4 cameras. MPI-INF-3DHP [50] annotates both 3D and 2D pose labels for human motion capture in a multi-camera studio. CMU Panoptic [35] presents a massively multiview system consisting of 31 HD Cameras to capture social interaction and provides 3D keypoints annotations of multiple people. HUMBI [76] collects local human behaviors such as gestures, facial expressions, and gaze movements from multiple cameras. AIST++ [41, 65] is a dance database that contains various 3D dance motions reconstructed from real dancers with multi-view videos. These datasets primarily focus on human activity motions ranging from daily activities to professional performances, rather than factors related to identity, cloth texture and body shape diversity. With the recent progress of neural rendering techniques, NHR [71], ZJU-Mocap [55], Neural Actor [24, 25, 44] and THuman4.0 [81] present their multi-view human dataset for evaluating the proposed human rendering algorithms. HuMMan [5] and Genebody [12] expand the diversity of pose actions and body shapes for human action recognition and modeling. ActorsHQ [30] uses dense multi-view capturing for photo-realistic novel view synthesis but is limited to 16 motion sequences and 8 actors. Recently, with the presence of the large-scale synthetic data and real captures from Objaverse [18, 19] and MVImgNet [75], several methods [29, 45] have made remarkable strides in the direction of open-world 3D reconstruction and generation. The concurrent work, DNA-Rendering [13] emphasizes the comprehensive benchmark functionality, but it encounters challenges in expanding the dataset to a larger scale due to the consideration of unusual human-object interactivity and clothes texture complexity. Differing from these efforts, we take a significant step forward in scaling up the human subjects and outfits, leading to the creation of MVHumanNet, the multi-view human capture dataset on the largest scale.

3. MVHumanNet

In this section, we provide a comprehensive overview of the core features of MVHumanNet, with a focus on dataset construction. We discuss the hardware capture system, data collection arrangements, dataset statistics, and data preprocessing. Sec. 3.1 provides an illustration to the fundamental aspects of the data acquisition system. This part specifically outlines the key components of the hardware capture system and its capabilities. Sec. 3.2 delves into

the actual data acquisition process, providing detailed information on personnel arrangement and the protocols followed during data collection. This section elucidates the steps taken to ensure the accuracy and consistency of the acquired data. Finally, in Sec. 3.3, we present a comprehensive framework that combines manual annotation and existing algorithms to obtain diverse and rich annotations for MVHumanNet. This framework enhances the applicability of our dataset for various research purposes.

3.1. Multiview Synchronized Capture System

We collected all the data using two sets of synchronized indoor video capture systems. In this section, we provide a detailed account of one system, while supplementary materials contain additional information about the second system. The primary framework of the capture system consists of 48 high-definition industrial cameras with a resolution of 12MP. These cameras are arranged in a multi-layer structure resembling a 16-sided prism, as shown in Fig. 1. The collection system has approximate dimensions of 2.4 meters in height and a diameter of roughly 4.5 meters. Each prism within the system is equipped with three 4K high-definition industrial cameras positioned at different heights.The lenses of each camera are meticulously aligned towards the center of the prism. To ensure clear image capture from different perspectives, we have placed light sources at the center of each edge of the system. During the data collection process, the frame rate of all cameras is set to 25 frames per second, enabling the capture of smooth and detailed motion sequences. For more comprehensive technical details, please refer to the supplementary materials.

3.2. Data Capture and Statistics

Data Capture To capture the wide range of dressing habits observed in people's daily lives, we establish a comprehensive process for performer recruitment and data collection. Specifically, at regular intervals, we release targeted recruitment requests to the public based on the statistics derived from the already collected clothing data. This strategy aims to enhance the diversity of clothing styles and colors for more reasonable human data distributions to achieve more reasonable human data distributions. In accordance with the clothing requirements, each performer is instructed to bring two sets of clothing to the capture system. Prior to the beginning of the capturing, performers randomly select 12 sets of actions from a predefined pool of 500 actions. Subsequently, they enter the capture system and sequentially perform the first six sets of actions, following instructions provided by the collection personnel. Each action is performed at least once on both the left and right sides for complete execution of the human performance capture. Upon completing the sixth set of actions, the performer finishes the first collection session by extending their hands to an A-pose
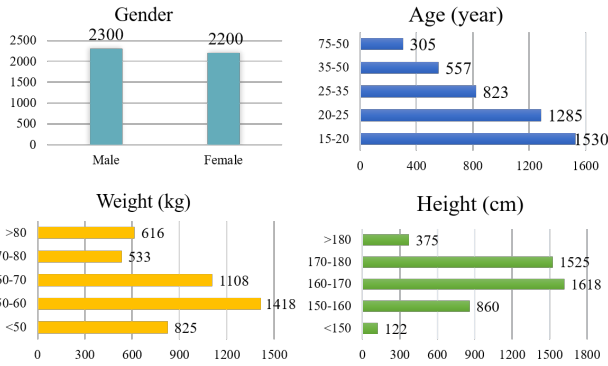
Figure 2. **The distribution of performers' attributes.** The gender, age, weight, and height of performers are recorded and carefully controlled. The statistical analysis of these attributes reflects a diverse range among the performers involved in MVHumanNet.
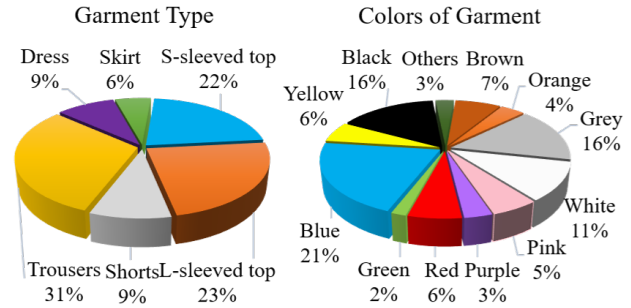

Figure 3. **The garment type and color distribution of outfits of performers.** Diverse colors and types of dressing are required for each invited performer. The statistical results show the wide coverage of daily clothes.

and rotating in place twice. Subsequently, the performer changes outfit and repeats the same process to complete the remaining six sets of actions with rotations in place.

Data Statistics The essential statistics of our dataset are shown in Fig. 2 and Fig. 3. MVHumanNet comprises a total of 4,500 unique identities with a equitable distribution of 2,300 male and 2,200 female individuals, ensuring a balanced representation of genders. Participants are required to fall within the age range of 15 to 75 years old. This age range is chosen to encompass a wide spectrum of performers while considering the potential impact of age on the quality and capabilities of their actions. Conversely, no restrictions are imposed on performers' weight or height, as these variables are deemed to have minimal impact on the data collection process. By not imposing such limitations, we aim to capture a more diverse and realistic representation of subjects in the dataset, allowing for a broader range of body types and proportions. Our dataset boasts the largest number of unique identities and garment items when compared to existing multi-view human dataset . It encompasses a wide range of everyday clothing styles and colors that are commonly available in real-world scenarios.

3.3. Data Annotation

To enable the advancement of applications in 2D/3D human understanding, reconstruction and generation, our dataset offers comprehensive and diverse annotations alongside the raw data. These annotations include action localization, attribute description, human masks, camera calibrations, 2D/3D skeleton, and parametric model fitting. The annotation pipeline, as depicted in Fig. 4, provides an overview of the entire process.

Manually Annotation Before capturing human data, we collect the cloth color and dress type of each performer in the registration table for further manual textual description. During the data collection process, we ensure a continuous flow as performers execute a sequence of six distinct ac-

tions along with in-place rotations. Subsequently, after the recording session, we manually mark the breakpoints for each action, accurately documenting the start and end of each action sequence. Moreover, the supplementary materials provide comprehensive records and annotations of the performers' basic attributes and outfits. For further details, please refer to the supplementary materials.

Human Mask Segmentation MVHumanNet comprises approximately 645 million images of individuals captured from various perspectives. Manual segmentation of such a massive image collection is obviously infeasible. To tackle this challenge, we propose a hierarchical automated image segmentation approach based on off-the-shelf segmentation algorithms. Our approach follows a coarse-to-fine segmentation strategy. Initially, we employ the RVM [42] to obtain efficient rough segmentation results. Subsequently, the rough segmentation outcomes are utilized to generate a bounding box of the performer, which serves as a prompt for the SAM [39] to produce higher-quality masks. In Fig. 4, the bottom-left region presents a comparison between the coarse and fine segmentation results.

Camera Calibration We utilized a commercial solution based on CharuCo boards to achieve fast and efficient camera calibration. Specifically, we position a CharuCo patterned calibration board at the central location of the capture studio. This ensures that each camera can capture a clear and complete view of the calibration board. With the aid of specific software, we obtain the intrinsic, extrinsic parameters and distortion coefficient for each camera. We also carefully adjust parameters, such as lighting, exposure, and camera white balance to capture high-quality data.

2D/3D Skeleton and Parametric Models Following the previous works [5, 12, 13] and with the goal of facilitating extensive research and applications in 3D digital human community, we conducted pre-processing on the entire dataset to obtain corresponding 2D/3D skeletons and two parameterized models. The processing pipeline is visually depicted in the bottom-right section of Fig. 4. Specifically, we employed the OpenPose [6] to predict 2D skeletons for
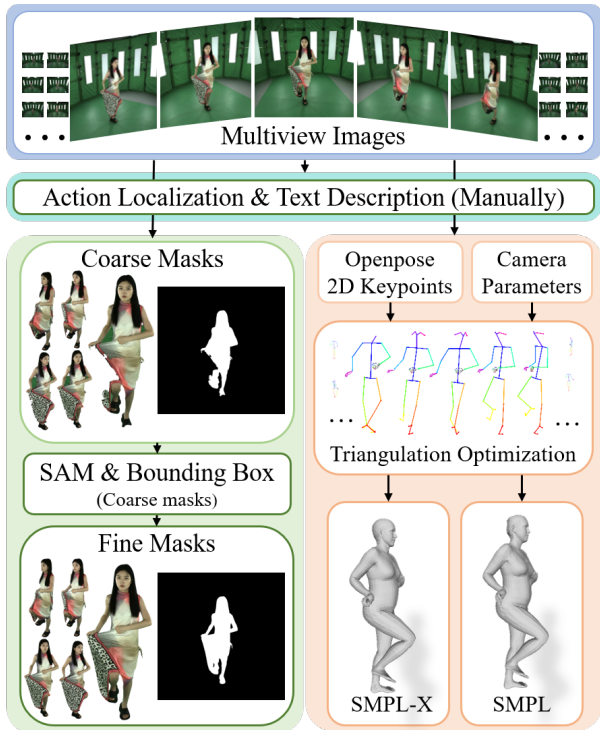
Figure 4. **Data annotation pipeline.** The manual and automatic annotation pipeline for action localization, text description, masks, 2D/3D keypoints and parametric models.

each frame of the images. Leveraging the calibrated camera parameters, multi-view 2D skeletons, and optimization algorithms [17], we derived the 3D skeletons from multi-view triangulation. Finally, SMPL/SMPLX parameters are fitted with the constrains of multi-view 2D keypoints and 3D skeletons. All these labeled data support MVHumanNet to be applied to various tasks.

4. Experiments

In this section, we present a comprehensive series of exploratory experiments conducted in the human action understanding, reconstruction, and generation tasks. Specifically, Sec. 4.1 focuses on showcasing experiments pertaining to view-consistent action recognition. As the dataset expands from single-view 2D data to multi-view 3D data, existing algorithms may encounter new challenges. In Sec. 4.2, we demonstrate experiments on generalizable NeRF (Neural Radiance Fields) reconstruction approaches, emphasizing the augmented model performance and generalization capabilities resulting from the increased availability of data. At last, in Sec. 4.3 and Sec. 4.4, we delve into recent research tasks, specifically text-driven view-unconstrained image generation and 3D human avatar generative model. Taking into account the size of the dataset, hardware limitations, and data annotation constraints, we performed experiments utilizing 62% of the available data.

	Train views	CTR-GCN[10]	InfoGCN[14]	FR-Head[82]
Top-1 (%)↑	1-view	33.85	25.23	30.25
	2-views	60.33	55.89	59.16
	4-views	72.16	73.59	71.74
	8-views	76.73	76.55	78.19
Top-5 (%)↑	1-view	51.08	37.14	50.59
	2-views	80.09	75.00	78.80
	4-views	88.32	89.02	88.67
	8-views	91.34	91.00	92.45

Table 2. Performance comparison of skeleton-based action recognition SOTA methods on MVHumanNet. With the increase of the views, the accuracy of the action prediction increases together.

More precisely, we employed 2800 identities, each representing a unique set of attire, amounting to a total of 5500 sets. Within this subset, 10% of the data was reserved exclusively for testing purposes.

4.1. View-consistent Action Recognition

MVHumanNet provides action labels with 2D/3D skeleton annotations, which can verify its usefulness on action recognition tasks. To simulate real-world scenarios, we employed single-view 2D skeletons as input and conducted tests on a multi-view test set that accurately represented real scenes. Our experimentation involved 8 viewpoints spaced at 45-degree intervals. The training data encompassed approximately 4000 outfits, while the testing data included 400 outfits, covering a total of 500 action labels. The results, presented in Tab. 2, reveal that the accuracy of action estimation was notably low for a single viewpoint, achieving a top-1 accuracy of only around 30%. However, as the number of input viewpoints increased, the accuracy of action estimation exhibited a significant improvement, peaking at 78.19%. Given that the dataset covers a comprehensive range of daily full-body actions, we possess confidence in its efficacy for facilitating diverse understanding tasks. Considering the challenges associated with acquiring 3D skeletons in everyday life, see supplementary for the results of 3D skeleton-based action recognition.

4.2. NeRF Reconstruction for Human

MVHumanNet can also be applied to NeRF reconstruction for human. Currently, human-centric methods, *e.g.* GP-NeRF [8], are developed in the context of lacking multi-view human data and their performance is still far from satisfactory on more diverse testing cases. We hope our proposed MVHumanNet can motivate more extensive studies of generalizable NeRF for human with sufficiently large scale of data. We empirically explore the performance of two distinct generalizable NeRFs methods, IBRNet [70] which is designed for general scenes and GPNeRF [8] which relies on human prior (*i.e.* SMPL [47]), using varying

Number of	IBRNet [70]			GPNeRF [8]		
outfits	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
100	26.05	0.9571	0.0555	23.27	0.8688	0.2077
2000	27.45	0.9638	0.0486	24.14	0.8779	0.2137
5000	29.00	0.9706	0.0377	24.69	0.8878	0.1961

Table 3. **Quantitative comparison of generalizable NeRFs with different scales of data for training.** We compare the results of methods with human prior and without human prior. We refer human prior to the commonly used SMPL model.
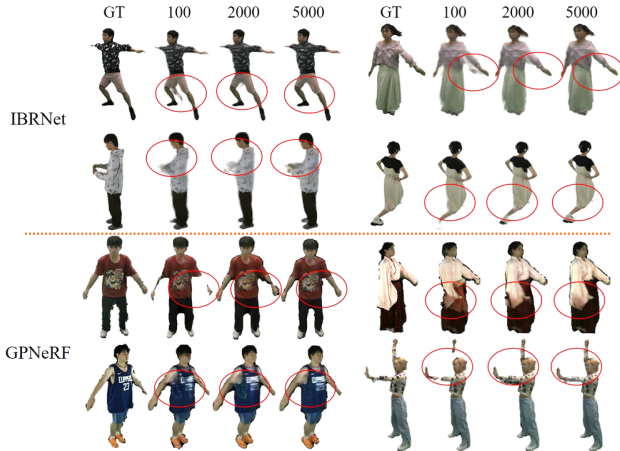
Figure 5. **The novel view synthesis results of IBRNet and GPN-eRF on unseen data of MVHumanNet. GT** means ground truth. The number of **100**, **2000**, and **5000** indicate the respective quantities of outfits utilized during the training process.

Method	IBRNet [70]			GPNeRF [8]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Train from scratch	28.06	0.9679	0.0437	20.95	0.9049	0.1809
w/o fintune	27.48	0.9663	0.0440	20.15	0.8921	0.2050
w/ fintune	29.46	0.9734	0.0323	21.89	0.9252	0.1364

Table 4. **Using MVHumanNet to pretrain a strong model**. We first train the representative methods on MVHumanNet, and then finetune the trained models on the train set of HuMMan [5]. We compare the performance of the finetuned models and models trained from scratch on the test set of HuMMan.

amounts of data for training. In our experiment, both approaches utilize four evenly distributed views as input and inference the novel view results. The quantitative comparisons of the outcomes are presented in Tab. 3, while the visualization results can be found in Fig. 5. Experimental results confirm that as the training data increases, the model exhibits enhanced generalization capabilities for new cases, especially when facing rare poses and complex garments. Moreover, we provided empirical evidence that MVHuman-Net can also serve for pretraining strong models, facilitating methods to perform better on out-of-domain scenarios. The corresponding results are presented in Tab. 4. Please note that the quantitative results of IBRNet [70] and GPN-eRF [8] cannot be directly compared, as they have different evaluation settings. More detailed explanations are in Supp.
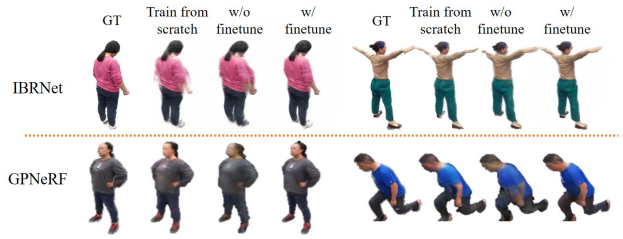

Figure 6. **Qualitative comparison of IBRNet and GPNeRF on the test set of HuMMan.** Without finetuning, the models only trained on MVHumanNet may suffer from domain gap. With some time for finetuning, the models outperform the ones trained merely on the train set of HuMMan.

4.3. Text-driven Image Generation

MVHumanNet is able to serves as a fundamental resource for our text-driven image generation method. The inclusion of comprehensive pose variations within our dataset enhances the potential for generating diverse human images in accordance with text descriptions. We finetune the powerful text-to-image model, Stable Diffusion [58] on MVHumanNet dataset to enable text-driven realistic human image generation. As shown in Fig. 7, given a text description and a target SMPL pose, we can produce high-quality results with the same consistency as text description and SMPL.

Based on the results derived from the text-driven image generation, it becomes evident that the utilization of large-scale multi-view data from real capture contributes to the efficacy of text-driven realistic human image generation.

4.4. Human Generative Model

Recently, generative models have become a prominent and highly researched area. Methods such as Style-GAN [21, 37] have emerged as leading approaches for generating 2D digital human. More recently, the introduction of GET3D [22] has expanded this research area to encompass the realm of 3D generation. With the availability of massive data in MVhumanNet, we embark on an exploratory journey as pioneers, aiming to investigate the potential applications of existing 2D and 3D generative models by leveraging a large-scale dataset comprising real-world data.

2D Generative Model Giving a latent code sampled from Guassian distribution, StyleGAN2 outputs a reasonable 2D images. In this part, we feed approximately 198,000 multi-view A-pose images (5500 outfits) and crop to 1024×1024 resolution into the network with camera conditions for training. Fig. 8 visualizes the results. Our model not only produces frontage full-body images but also demonstrates the capability to generate results from other views, including the back and side views.

3D Generative Model Unlike StyleGAN2, GET3D [22] introduces a distinct requirement of one latent code for geometry and another for texture. We use the same amount of data as training StyleGAN2 to train GET3D. The visual-
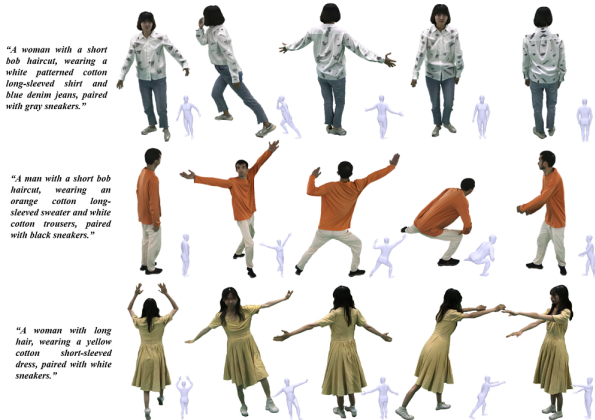
Figure 7. **The visualization of images generated by text-to-image model trained on MVHumanNet with SMPL condition and text prompts as input.** The results demonstrate that training on our large-scale high-quality human dataset enables the generation of high-resolution human images using textual description and SMPL conditions. Supp. shows more results.

Figure 8. **Visualize the results of StyleGAN2 trained with MVHumanNet.** We randomly sample latent codes from Gaussian distribution and obtain the results. See supp. for more results.

Number of Subjects	FID↓	
	StyleGAN2 [37]	GET3D [22]
3000	14.05	41.54
5500	7.08 (-6.97)	25.12 (-16.42)

Table 5. **Quantitative comparison of generative models with different data scale**. The performance of both 2D and 3D generative models exhibits obvious improvement with scaling up data.

ization results are shown in Fig. 9. The model exhibits the ability to generate reasonable geometry and texture in the A-pose, thereby enabling its application in various downstream tasks. With the substantial support provided by MVHumanNet, various fields, including 3D human generation, can embark on further exploration by transitioning from the use of synthetic data or single-view images to the incorporation of authentic multi-view data. We also conduct experiments to prove that the performance of the generative model will become more powerful with the increase in the amount of data. The quantitative results are shown in Tab. 5. We have reason to believe that with the further increase of data, the ability of trained models can further improve.
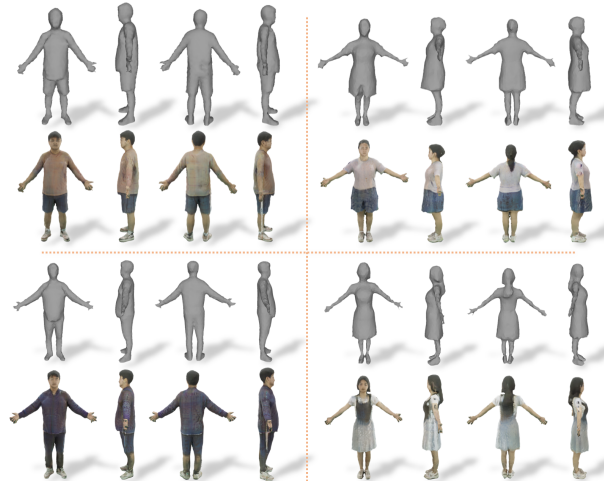

Figure 9. **The visualization results of GET3D trained with MVHumanNet rendered by Blender [16].** The first and third rows represent the geometry, while the second and fourth row shows the texture corresponding to geometry.

5. Conclusion

In this work, we present MVHumanNet, a large-scale multi-view dataset containing **4,500** human identities, **9,000** daily outfits and **645 million** frames with extensive annotations. We primarily focus on the domain of collecting daily dressing, which allows us to easily scale up the human data. To probe the potential of the proposed large-scale dataset, we design four experiments to show how MVHumanNet can be used to power these 3D human-centric tasks. We plan to release the MVHumanNet dataset with annotations publicly and hope that it will serve as a foundation for further research in the 3D digital human community. To mitigate potential negative social impacts, we will implement strict regulations on the utilization of our data.

Acknowledgement

References

[1] 3d people. https://3dpeople.com/. 3

[2] AXYZ. https://secure.axyz-design.com/. 3

[3] Federica Bogo, Javier Romero, Matthew Loper, and Michael J Black. Faust: Dataset and evaluation for 3d mesh registration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014. 3

[4] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[5] 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. In *European Conference on Computer Vision*. Springer, 2022. 3, 4, 5, 7

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 5

[7] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[8] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *ECCV*, pages 222–239. Springer, 2022. 6, 7

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 3

[10] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*, 2021. 6

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

[12] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 2, 3, 4, 5

[13] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *ICCV*, 2023. 2, 3, 4, 5

[14] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *CVPR*, 2022. 6

[15] CMU Graphics Lab. http://mocap.cs.cmu.edu/. 3

[16] Blender Online Community. Blender - a 3d modelling and rendering package, 2018. http://www.blender.org. 8

[17] EasyMocap Contributors. Easymocap - make human motion capture easier. Github, 2021. https://github.com/zju3dv/EasyMocap. 6

[18] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2, 4

[19] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 4

[20] Anna Frühstück, Krishna Kumar Singh, Eli Shechtman, Niloy J Mitra, Peter Wonka, and Jingwan Lu. Insetgan for full-body image generation. In *CVPR*, 2022. 3

[21] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*. Springer, 2022. 3, 7

[22] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances In Neural Information Processing Systems*, 35, 2022. 3, 7, 8

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 3

[24] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 4

[25] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (ToG)*, 40(4), 2021. 4

[26] 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 *CVPR*, 2023. 3

[27] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3

[28] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 3

[29] 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, 2023. 2, 4

[30] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias

Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2, 3, 4

[31] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 3, 4

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 3

[33] Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, and Lan Xu. Humangen: Generating human radiance fields with explicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[34] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 3

[35] 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 *ICCV*, 2015. 3, 4

[36] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[37] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 7, 8

[38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 2

[39] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, pages 4015–4026, 2023. 5

[40] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 2021. 3

[41] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 2, 3, 4

[42] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, pages 238–247, 2022. 5

[43] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[44] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)*, 40(6), 2021. 3, 4

[45] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023. 2, 4

[46] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 3

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics(ToG)*, 34(6), 2015. 6

[48] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *CVPR*, 2020. 3

[49] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, 2019. 3

[50] 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. In *International conference on 3D vision (3DV)*. IEEE, 2017. 3, 4

[51] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 3

[52] 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. 3

[53] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4), 2022. 3

[54] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 3

[55] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3, 4

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[57] Renderpeople. `https://renderpeople.com/`. 2, 3

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 7

[59] Huang Geng Sai Charan Mahadevan, Karunanidhi Durai Kumar. `https://mocap.cs.sfu.ca/`. 3

[60] 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 *ICCV*, 2019. 3

[61] 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. 3

[62] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35, 2022. 2

[63] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016. 3

[64] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *CVPR*, 2023. 3

[65] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, 2019. 4

[66] Twindom. `https://web.twindom.com/`. 3

[67] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics (TOG)*, 27(3):1–9, 2008. 4

[68] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009. 4

[69] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *CVPR*, 2014. 3

[70] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 6, 7

[71] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *CVPR*, 2020. 3, 4

[72] Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, and Xiaoguang Han. Get3dhuman: Lifting stylegan-human into a 3d generative model using pixel-aligned reconstruction priors. In *ICCV*, 2023. 3

[73] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *CVPR*, 2022. 3

[74] 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 *CVPR*, pages 5746–5756, 2021. 2, 3

[75] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 2, 4

[76] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *CVPR*, 2020. 3, 4

[77] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 3

[78] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017. 3

[79] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *ICCV*, 2021. 3

[80] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[81] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022. 3, 4

[82] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *ICCV*, 2023. 6