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

In the appendix, we provide detailed information about the proposed MVHumanNet and various experiments conducted on our dataset. Sec. 1 introduces the other capture system and provides details about the data collection and annotation. Sec. 2 visualizes additional experimental results.

1. Dataset Details

1.1. The Second Multiview Capture System

We utilize two sets of synchronized indoor video capture systems to collect the MVHumanNet dataset. We have provided a detailed account of one system in the main text, while we introduce the second system here. The second capture system consists of 24 high-definition industrial cameras which are evenly distributed on 16 pillars in a two-layer structure, as shown in Fig. 1. The collection system has approximate dimensions of 2.2 meters in height and roughly 4.3 meters in diameter. The lenses of each camera are meticulously aligned towards the center of the prism. To ensure clear image capture from different perspectives, we place 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 30 frames per second, enabling the capture of smooth and detailed motion sequences.

We capture a total of 9,000 outfits by using these two sets of systems. The 48-view system captures approximately 5,000 outfits with a resolution of 4096×3000, while the second system accounts for the remaining 4,000 outfits with a resolution of 2448×2048.

Camera Calibration We utilize the same commercial solution based on CharuCo boards to achieve fast and efficient camera calibration. Recognizing the potential for performers to inadvertently come into contact with the capture studio or cameras during their entry or execution of actions, we implement a calibration process at the beginning, middle, and end of each day. This procedure aims to account for any potential changes in camera parameters. We also carefully adjust other parameters, such as lighting, exposure, and camera white balance to capture high-quality data.

1.2. Action Statistics

We also make efforts to prepare diverse performed actions that cover a broad spectrum of action categories, including sports, social engagements, education, entertainment and professional actions, as shown in Fig. 5. These categories collectively contribute to the incorporation of 500 distinct actions within our MVHumanNet dataset, providing a comprehensive range of options.

1.3. Coarse to Fine Masks

Based on the RVM [6] and powerful SAM [5], we propose a coarse-to-fine mask segmentation strategy. Specifically, we leverage RVM to propose a coarse mask and a candidate bounding box for the performer. Subsequently, the candidate bounding box serves as input to the robust SAM, enabling the prediction of refined masks. The comparative results between the coarse and fine masks are visually demonstrated in Fig. 2. From left to right, the image illustrates the input image, coarse mask, fine mask, and

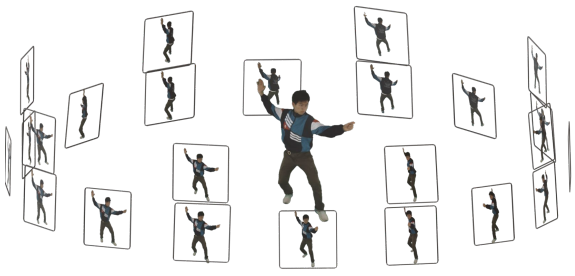

Figure 1. **The visualization of the second multiview synchronized capture system.** Our second capture system consists of 24 industrial cameras with a resolution of 2448×2048.

Figure 2. **The visualization results of our coarse to fine mask segmentation strategy.** Note that we crop all images into a square to show the comparison results.

segmented performer. The results clearly indicate that the coarse mask exhibits more substantial errors. For instance, in the first row, the coarse mask fails to accurately segment the background area between the arm and the body, while the fine mask successfully addresses this issue.

1.4. Text Description

Throughout the entire process of data collection, we carefully record the essential details of each identification encompassing crucial information such as gender and age. Furthermore, we employ manual labeling to furnish text descriptions of the performers' hairstyles and shoes, as well as each outfit, including clothing color, style and material. Fig. 3 provides a visual representation, while Fig. 4 offers additional examples. These text descriptions can be utilized to support tasks such as text-to-image generation, as demonstrated in the experimental results presented in Sec. 2.2.

2. Experimental Results

In Sec. 1.1, we illustrate our utilization of two capture systems to collect a total of 9,000 sets of clothing. Throughout all experiments, we employ data from both capture systems in a 1:1 ratio. For instance, in Sec. 2.1, we train NeRFs using a maximum of 5,000 outfits, with each capture system contributing approximately half of the data.

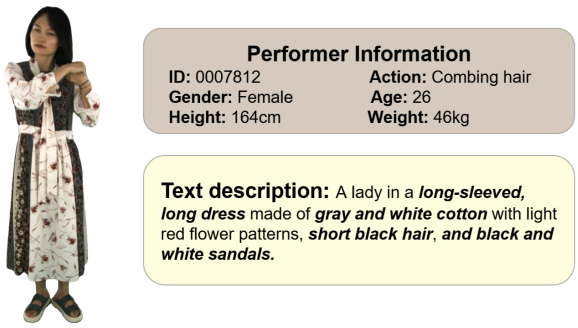

Figure 3. **A text description demo.** The description contains various information, such as age, height, garment and hairstyle.

2.1. NeRF Reconstruction for Human

In the main text, we have presented comprehensive objective and subjective comparison results of training a generalizable NeRF model using varying amounts of data, along with fine-tuning experiments on HuMMan [1] using MVHumanNet. We include additional visualizations here. Fig. 6 showcases the comparative results of training IBRNet [8] with 100, 2,000, and 5,000 outfits, while Fig. 7 demonstrates the corresponding outcomes of training GPNeRF [2] with the same datasets. Additionally, Fig. 8 further demonstrates the comparison results of fine-tuning on HuMMan [1] using MVHumanNet as a pretraining dataset. The supplementary video materials contain a more extensive collection of visual results.

Differences in evaluation settings between IBRNet and GPNeRF. As GPNeRF [2] is specially designed for human rendering, it exploits the projected SMPL model to crop the human area of a rendered image and only evaluates the human area with the three metrics. In contrast, IBRNet [8] is a generalizable NeRF method for general scenes, so it evaluates the whole rendered image by default. Additionally, GPNeRF [2] masks the background area with black, while IBRNet [8] uses a white background.

2.2. Text-driven Image Generation

We utilize MVHumanNet dataset to finetune the Stable Diffusion [7], and the Fig. 9 visualize the extra results of text-driven image generation. Given a paragraph of **text description** and a **pose** as inputs, we can easily obtain the desired realistic results. For example, in the third row, when the input text contains keywords indicating a **ponytail**, the output images align well with the expected results.

2.3. Human Generative Model

Additional results of **2D** and **3D** generative models are presented in Fig. 10 and Fig. 11. Based on the obtained results, it is evident that the visual quality of the 2D output

outperforms that of the 3D results, which is consistent with our initial expectations. This disparity can be attributed to the relatively early stage of 3D research compared to the well-established domain of 2D digital human generation. One significant contributing factor to this discrepancy is the limited availability of 3D datasets. We hope that with the opening of MVHumanNet, it can help the community further explore the relevant methods.

Performer Information
ID: 0003562 **Action:** Throwing rope
Gender: Male **Age:** 18
Height: 176cm **Weight:** 60kg

Text Description: A man dressed in ***Han-style gray polyester top*** with printed dragons, ***beige nylon trousers***, ***short black hair***, and ***gray and white sneakers***.

Performer Information
ID: 0008230 **Action:** Horizontal stride
Gender: Female **Age:** 28
Height: 159cm **Weight:** 52kg

Text Description: A woman wearing a ***pure red cotton sleeveless shirt*** and ***long skirt***, with ***short black hair***, ***barefoot***.

Performer Information
ID: 0008020 **Action:** Cross-legged meditation
Gender: Male **Age:** 22
Height: 175cm **Weight:** 65kg

Text Description: A man in a ***white, short-sleeved, mixed fabric T-shirt*** with lapels and letters, ***long pants***, ***short black hair***, and ***gray and white sneakers***.

Performer Information
ID: 0000105 **Action:** Greeting with one hand
Gender: Female **Age:** 57
Height: 170cm **Weight:** 59kg

Text Description: A woman ***with long-straight brown hair*** wearing a ***long-sleeved white polyester shirt*** and ***blue chiffon trousers***, with ***gray sneakers***.

Performer Information
ID: 0006065 **Action:** Watching the time
Gender: Male **Age:** 19
Height: 175cm **Weight:** 62kg

Text Description: A man wearing a ***khaki polyester suede coat***, ***cyan denim jeans***, ***short black hair***, and ***black and white sneakers***.

Performer Information
ID: 0008728 **Action:** Sitting with arms crossed
Gender: Female **Age:** 20
Height: 165cm **Weight:** 55kg

Text Description: A lady wearing a ***multicolored patterned cotton long-sleeved shirt*** over a ***white cotton V-neck vest***, with ***black cotton hot pants***, ***long black hair***, and ***white leather shoes***.

Performer Information
ID: 0008021 **Action:** Provocation
Gender: Male **Age:** 23
Height: 182cm **Weight:** 65kg

Text Description: A man wearing a ***red and blue spider-man pattern cotton hoody***, ***brown cotton trousers***, ***short black hair***, and ***black and white sneakers with orange shoelaces***.

Performer Information
ID: 0001122 **Action:** Rock-climbing
Gender: Female **Age:** 35
Height: 165cm **Weight:** 45kg

Text Description: A lady wearing a ***yellow and white striped V-neck knit cardigan***, ***orange chiffon long skirt***, ***short black hair***, and ***black sandals***.

Figure 4. **Sampled examples of text descriptions for different performers.**

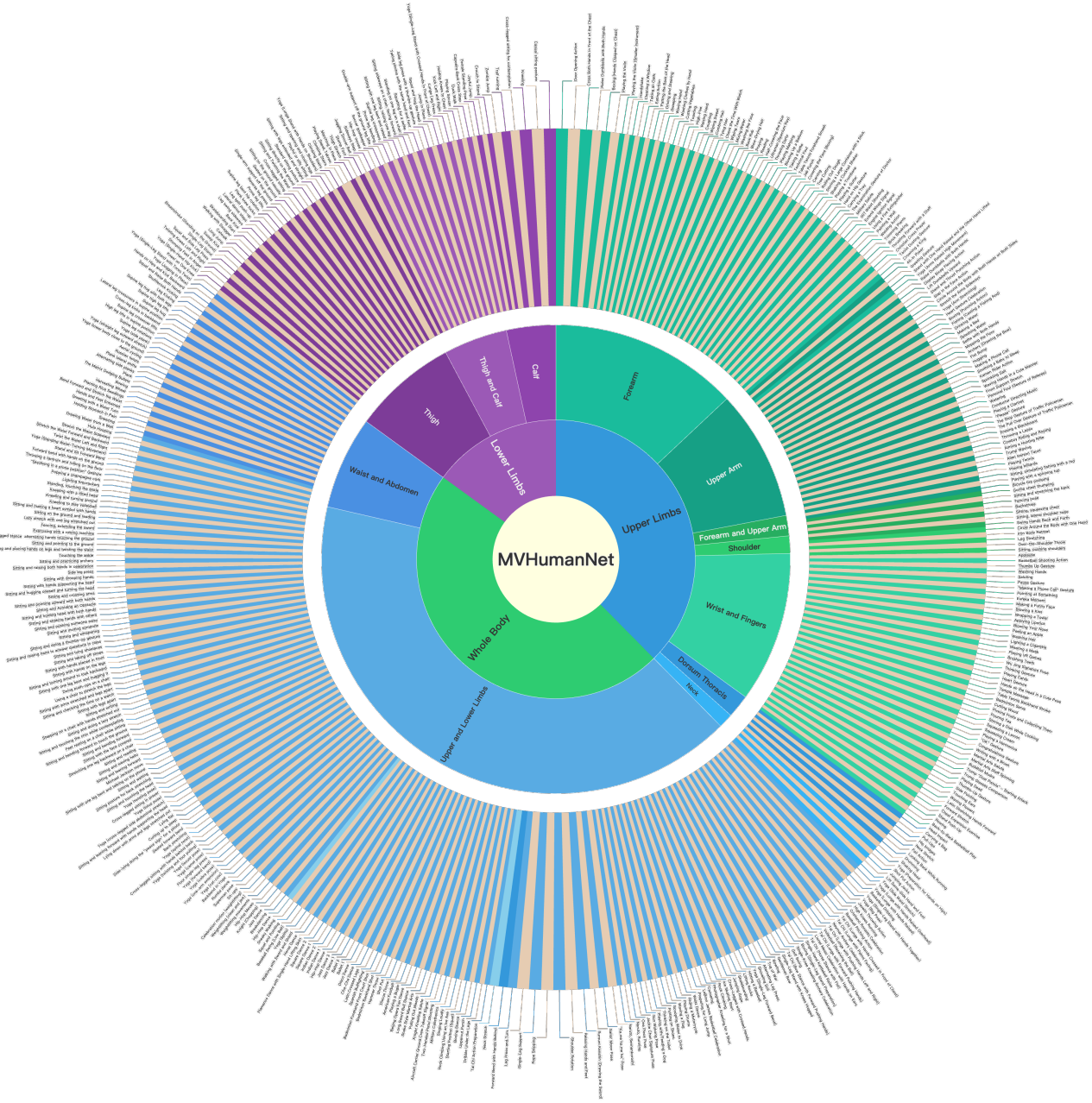

Figure 5. **The visualization of 500 action labels.**

Figure 6. **More visualization results of the IBRNet [8]. GT** means ground truth. The number of **100**, **2,000**, and **5,000** indicate the respective quantities of outfits utilized during the training process.

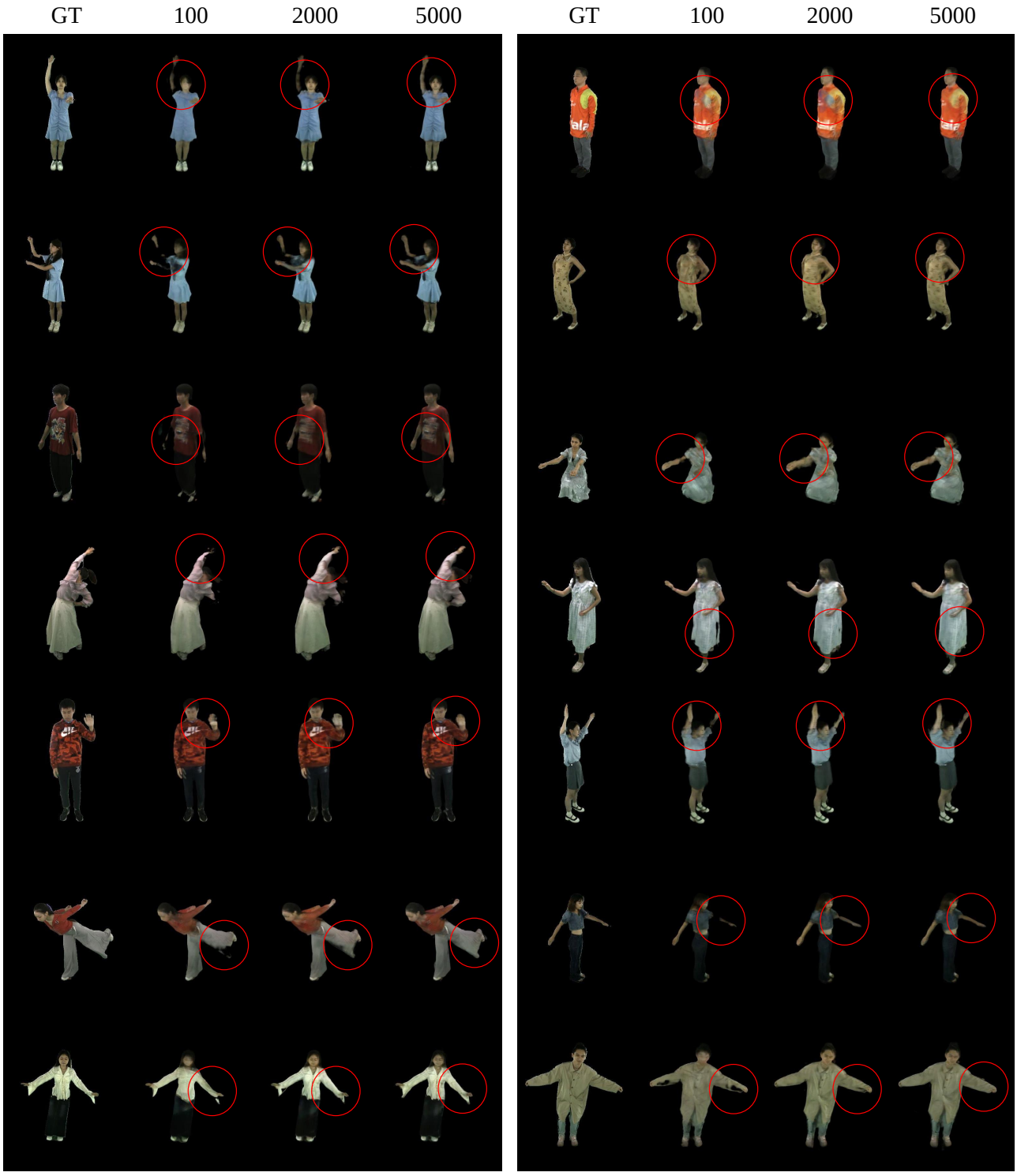

Figure 7. **More visualization results of the GPNeRF [2]. GT** means ground truth. The number of **100**, **2,000**, and **5,000** indicate the respective quantities of outfits utilized during the training process.

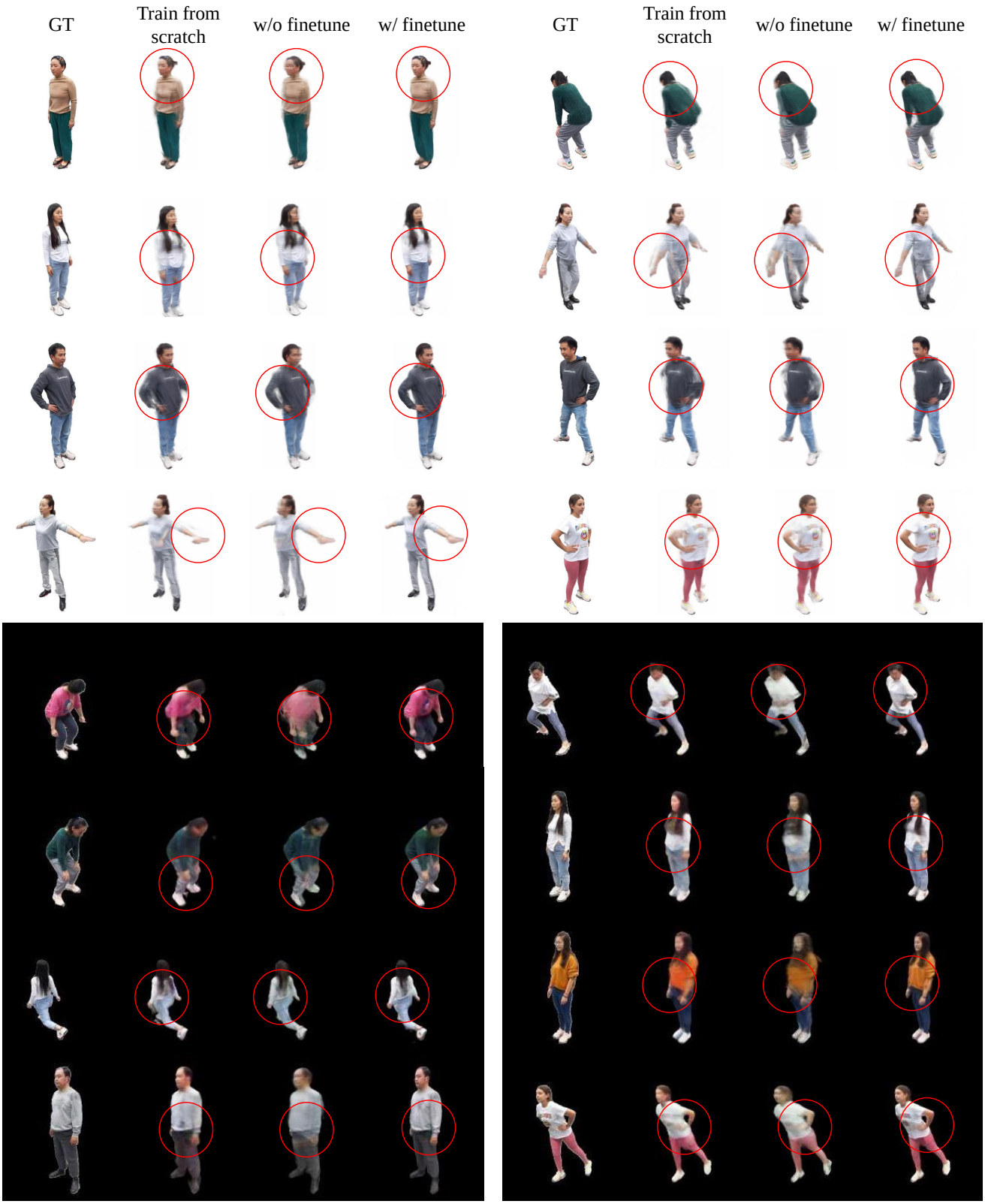

Figure 8. **More visualization results of IBRNet [8] (top) and GPNeRF[2] (bottom) fine-tuned on HuMMan [1] dataset.**

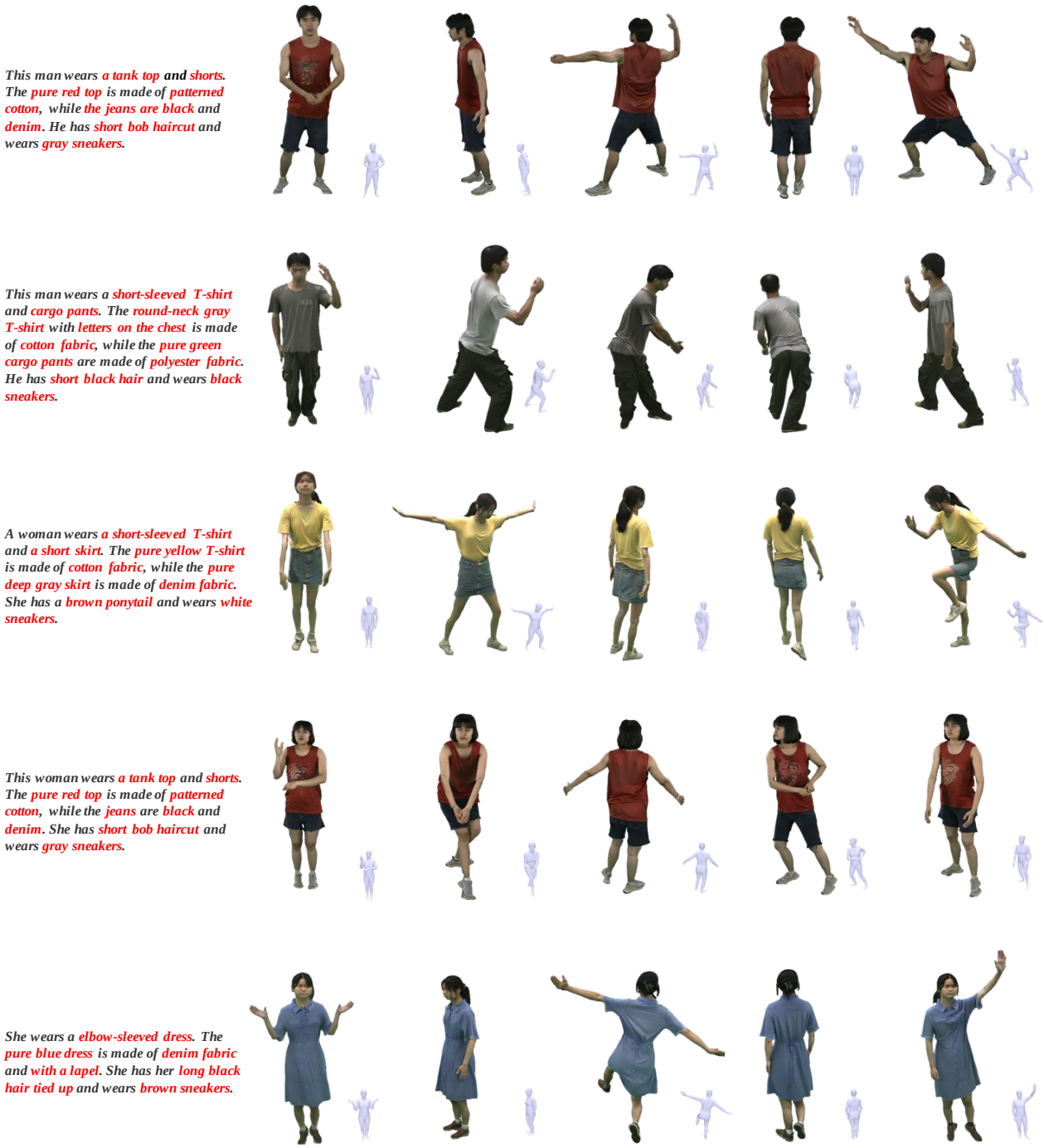

Figure 9. **More visualization results of text-driven image generation.** Given a text description and a target pose, we can produce high-quality results with the same consistency as text description and SMPL.

Figure 10. **More visualization results of StyleGAN2 [4] trained with A-posed multi-view images in MVHumanNet.** We randomly sample latent codes from Gaussian distribution and obtain the results.

Figure 11. **More visualization results of GET3D [3] trained with A-posed multi-view images in MVHumanNet.**

References

[1] 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. 2, 8

[2] 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. 2, 7, 8

[3] 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. 11

[4] 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. 10

[5] 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. 1

[6] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *WACV*, pages 238–247, 2022. 1

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

[8] 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. 2, 6, 8