

---

# Supplementary Materials for GeneMAN: Generalizable Single-Image 3D Human Reconstruction from Multi-Source Human Data

---

Anonymous Author(s)

Affiliation

Address

email

1 This appendix serves to further enrich the discourse established in our main paper. Sec. A offers  
2 additional information to supplement the related work. Sec. B presents the implementation details of  
3 GeneMAN. In Sec. C, we elaborate on the details of our multi-source human dataset. Sec. D provides  
4 supplementary experimental results, including more visualizations for the qualitative comparison,  
5 additional results of GeneMAN on in-the-wild images and CAPE [24], discussion on geometric  
6 quantitative evaluation, and ablation studies conducted on our multi-source human dataset. In Sec. E  
7 and F, we discuss the limitations of our approach, as well as the broader impacts and safeguards.

## 8 A Supplement to Related Work

### 9 A.1 Text-to-3D Generation

10 The rapid advancements in 2D image generation have also significantly accelerated progress in the  
11 field of text-to-3D generation. DreamFusion [29] and Magic3D [16] have demonstrated that utilizing  
12 pretrained 2D text-to-image diffusion models [32, 31] as guidance can greatly enhance the optimiza-  
13 tion of 3D representations through score distillation sampling (SDS). Trained on large-scale 2D  
14 image datasets, these text-to-image models possess excellent capabilities in providing comprehensive  
15 2D knowledge and hallucinating unseen scenes based on specific prompts. Nevertheless, as these  
16 models are limited to 2D knowledge, the aforementioned text-to-3D methods are susceptible to  
17 multi-view consistency issues. MVDream [39] improves 3D consistency by training a multi-view  
18 diffusion model that offers a robust multi-view prior. However, these methods are not well-suited  
19 for direct image-to-3D reconstruction tasks because images cannot be accurately captured through  
20 textual descriptions, leading to inconsistencies in color and texture across the generated assets.

### 21 A.2 Single Image-to-3D Reconstruction

22 Single image-to-3D reconstruction has greatly benefited from the thriving advancement of diffusion  
23 models. Zero-1-to-3 [19] trains a view-dependent diffusion model on Objaverse [6] by explicitly in-  
24 corporating the camera parameters into the diffusion process. It enables zero-shot image-conditioned  
25 novel view synthesis and facilitates general image-to-3D reconstruction with consistent geometric  
26 view. Magic123 [30] and DreamCraft3D [42] integrate the 3D prior from Zero-1-to-3 [19] with  
27 2D prior from text-to-image diffusion model, employing a coarse-to-fine two-stage optimization to  
28 achieve high-quality 3D reconstruction. LRM [10] leverages both synthetic data from Objaverse [6]  
29 and multi-view data from MVImageNet [48] to train a transformer-based 3D architecture, enhancing  
30 its generalizability for 3D reconstruction. However, these general image-to-3D approaches often  
31 yield poor results in human reconstruction, resulting in inaccurate geometry of human bodies and the  
32 neglect of intricate details, such as facial features and clothing. This limitation arises from the lack of  
33 human-specific priors.

## 34 B Implementation Details

### 35 B.1 Camera Sampling

36 For novel view guidance, we sample the camera distance from the range  $\mathcal{U}(3.0, 3.5)$ . The elevation angle  $\phi$  is drawn from  $\mathcal{U}(-10^\circ, 45^\circ)$ , and the azimuth angle  $\theta$  is uniformly sampled from  $[-180^\circ, 180^\circ]$ . Additionally, the field of view (FOV) is constrained to the range  $[20^\circ, 25^\circ]$ , aligning with the fixed camera intrinsics optimized for our finetuned human diffusion models. Building upon HumanNorm [11], we segment the human body into four regions: the head, upper body, lower body, and full body. To enhance part-aware reconstruction with high-fidelity detail, we allocate a sampling probability of 0.7 to the full body, and 0.1 to each of the head, upper body, and lower body. We also manually adjust the camera distance for zooming in on specific parts. For instance, when refining the facial region, the camera distance is reduced to  $\mathcal{U}(0.8, 1.0)$ . For half-body refinement (either upper or lower body), the camera distance is adjusted to  $\mathcal{U}(1.5, 2.0)$ . Additionally, the camera center is shifted to ensure that the relevant keypoint aligns with the center of the rendered images. To achieve this, we use an off-the-shelf tool to identify 2D keypoints, which are then used during rasterization to back-project the coordinates into 3D space. Note that if a keypoint lies outside the image boundaries, we assign a sampling probability of zero to that point and renormalize the distribution accordingly.

### 50 B.2 Details of Each Stage

51 **Geometry Initialization & Sculpting.** In the phase of Geometry Initialization, we optimize Instant-NGP [26] from a resolution of 128 to 384 over the course of 5,000 steps. The loss weights for this stage are set as follows:  $\lambda_r = 1 \times 10^3$ ,  $\lambda_m = 100$ ,  $\lambda_d = 0.05$ ,  $\lambda_n = 1$ ,  $\lambda_{2D} = 0.1$ ,  $\lambda_{3D} = 0.1$ . Subsequently, we extract the resulting mesh from Instant-NGP with an isosurface resolution of 256 as the geometry initialization. During the geometry sculpting stage, we refine the geometry adopting DM Tet [38] at a resolution of 512 for 3,000 steps, enabling the capture of intricate details of humans. Inspired by HumanNorm [11], we incorporate a progressive positional encoding technique, where the mask on the position encoding for DM Tet’s SDF features is gradually lifted to introduce higher-frequency components as training progresses. After 2,000 iterations, the mask fully reveals all positions, allowing the encoding to capture both low- and high-frequency details. Empirically, we observe that omitting progressive positional encoding results in noisy surface reconstructions. The loss weights are set as follows:  $\lambda_r = 5 \times 10^3$ ,  $\lambda_{vgg} = 1 \times 10^3$ ,  $\lambda_{sdf} = 1.5 \times 10^3$  and  $\lambda_{sds} = 1.0$ . To ensure consistency between the rendered front view and the input image, we apply a relatively small guidance scale of 20 for novel view generation using the normal- and depth-adapted diffusion models [11]. The noise timestep  $t$  is sampled from  $\mathcal{U}(0.02, 0.8)$ . Besides, we adopt the AdamW [22] optimizer with a base learning rate of 0.01 during Geometry Initialization and  $2 \times 10^{-5}$  during Geometry Sculpting.

68 **Multi-Space Texture Refinement.** For latent space texture refinement, we employ SDS optimization [29] in the latent space to optimize the coarse texture at an image resolution of 1024  $\times$  1024 for 10,000 steps. The loss weights for the coarse texture stage are configured as follows:  $\lambda_{rgb}^{color} = 1 \times 10^3$ ,  $\lambda_m^{color} = 100$ ,  $\lambda_{2D}^{color} = 0.1$ ,  $\lambda_{3D}^{color} = 0.1$ . Following this, we perform pixel-space texture refinement by optimizing the UV texture map for an additional 1,000 steps to achieve finer texture details. To ensure that the added noise does not corrupt the original content while retaining the capacity to enhance image details, we empirically set the starting timestep  $t_{start}$  to 0.05 in Eq.8 in the main text. The weight for the VGG loss [40] in the pixel space texture refinement is set to  $\lambda_{LP} = 0.01$ . We adopt the AdamW [22] optimizer with a base learning rate of 0.01 in the coarse texture stage and 0.02 in the fine texture stage.

### 78 C Details of Multi-Source Human Dataset

79 To facilitate the training of multi-source human diffusion models, we construct a comprehensive multi-source human dataset containing 100K 2D human images and 52,345 multi-view 3D human instances in total. Detailed statistics of our multi-source human dataset are provided in Tab. 1, with its composition described as follows: (1) The 2D human data consists of 20K human images from DeepFashion [20] and 80K human images filtered from LAION-5B [35]. The filtering process is conducted using YOLOv7 [43], eliminating images without human subjects or containing multiple individuals. Additionally, images are excluded if they have an aesthetics score below 4.5, if the largest

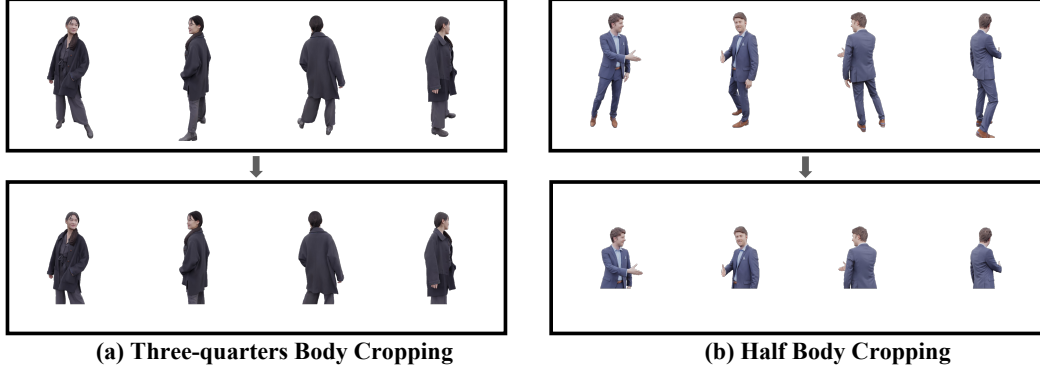


Figure 1: **Image Cropping Synthesis.** We apply half-body and three-quarters-length cropping to the multi-view renderings of 3D scans to generate synthetic cropped images.

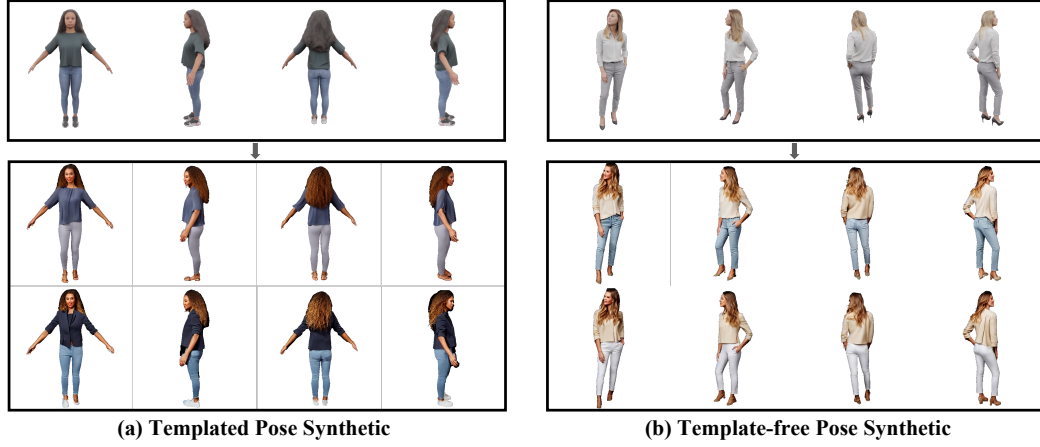


Figure 2: **ControlNet-based Synthesis.** We perform ControlNet-based synthesis for both templated poses and template-free poses.

86 detected face measures less than  $224 \times 224$  pixels, or if the overall resolution falls below  $640 \times 1280$ .  
87 To enhance semantic richness, each image is captioned using a finetuned BLIP model [15]. (2) The  
88 3D human data is collected from 3D scanned human data, synthetic human data, video human data,  
89 and our synthetic human data. **3D Scanned Human Data** contains human models sourced from  
90 the commercial dataset RenderPeople [1] and open-source datasets including CustomHumans [7],  
91 HuMMan [4], THuman2.0 [47], THuman3.0 [41] and X-Humans [37]. **3D Synthetic human data**  
92 contains human-category objects filtered from Objaverse [6]. For both scanned and synthetic data, we  
93 adopt the dataset creation protocol of Zero-1-to-3 [19], with the exception that we uniformly select 48  
94 viewpoints across 360 degrees in azimuth and set the resolution of the rendered image to  $1024 \times 1024$ .  
95 **Video Human Data** contains five open-source datasets: DNA-Rendering [5], ZJU-MoCap [28],  
96 Neural Actor [18], AIST++ [14] and Actors-HQ [13]. For datasets containing background imagery,  
97 we apply CarveKit [36] to extract human silhouettes and produce corresponding RGBA images.  
98 When datasets feature multiple groups of surrounding cameras, we prioritize the group capturing  
99 the human subject at the image center. Note that we filter out cases of matting failures in these  
100 video datasets. As for DNA-Rendering [5], only Parts 1 and 2 of the released data are utilized.  
101 **Our Synthetic Human Data** contains human data synthesized by two augmentation techniques:  
102 ControlNet-based [49] image synthesis and image cropping synthesis. To further enhance the diversity  
103 of human identities and outfits in the dataset, inspired by En3D [25], we use ControlNet to synthesize  
104 a batch of multi-view human data based on diverse prompts describing outfits, genders, ages, and  
105 more, all generated by ChatGPT [2]. Unlike En3D, we extract templated or template-free poses and  
106 depth images from multi-view renderings of scanned human data (RenderPeople and XHumans [37])  
107 as conditions fed into ControlNet and synthesize multi-view human images using diverse prompts. To  
108 ensure the quality of synthetic images, we concatenate pose and depth images horizontally across all

Table 1: **The Statistics of Multi-Source Human Dataset.** Our multi-source human dataset encompasses a diverse range of human data, categorized into 3D part data: 3D scans, multi-view videos, our synthetic data, and 2D data: single images. The number of 2D images and 3D instances for each dataset is summarized in the table below.

Multi-Source Human Dataset			
3D Part Data	3D Scanned and Synthetic Human Data		
	RenderPeople [1]	Thuman2.0 [47]	Thuman3.0 [41] HuMMan [4]
	1853	526	458 9072
	CostomHuman [7]	X-Humans [37]	Objaverse Human [6]
	647	3384	3388
	Video Human Data		
	DNA Rendering [5]	ZJU-Mocap [28]	Neural Actor [18]
	8780	2646	4800
	AIST++ [14]	Actors-HQ [13]	
	3853	3600	
	Our Synthetic Human Data		
	ControlNet-based Synthetic	Image Cropping Synthetic	
	5642	3706	
	2D Part Data		
	DeepFashion [20]	LAION-5B [35]	
	20K	80K	

views, generating multi-view humans in a single inference. This concatenation strategy ensures the texture consistency of the synthetic multi-view images. To handle the reconstruction of in-the-wild images with arbitrary human proportions, we crop the above fully body-length multi-view 3D human data into half-body length or three-quarters-body length images to create the augmented images. As for 3D data captioning, inspired by 3DTopia [9] and Cap3D [23], we use multi-modal large language model LLaVA [17] to generate captions for 3D objects by aggregating the descriptions from multiple views. The success of our GeneMAN framework demonstrates the effectiveness of our dataset in providing generalized priors for diverse human geometry and textures.

## D Additional Experiment Results

### D.1 More Qualitative Results

For a more comprehensive geometric evaluation, we incorporated six additional SOTA human geometry reconstruction methods for comparison: PIFuHD [34], PaMIR [52], ICON [46], ECON [45], PHORHUM [3], and SIFU [51]. The geometric comparison results are presented in Fig. 7 and Fig. 8. It can be seen that our model effectively recovers natural human poses, accommodates loose clothing, and excels at reconstructing images with diverse body ratios. In addition, we provide more qualitative comparisons with state-of-the-art methods: PIFu [33], GTA [50], TeCH [12], SiTH [8] as shown in Fig. 9 to Fig. 15. Our method surpasses the compared methods in generating consistent, highly realistic textures with exceptional fidelity.

Besides, in Fig. 3, we conduct additional comparisons with two generalizable methods: HumanLRM [44] and IDOL [53], both of which are feed-forward models trained on large-scale human datasets. As the code of Human-LRM is unavailable, we use the examples provided in its paper. Our model authentically reconstructs 3D humans that closely align with the front view, while IDOL fails to preserve facial details. Our model also produces natural poses with reasonable geometry (see the legs of the first and the third subjects in Fig. 3). Moreover, it achieves high-fidelity appearance with view-consistent details, significantly outperforming HumanLRM.

We provide additional visualization results, with a particular focus on complex poses, in Fig. 5 (in-the-wild images) and Fig. 6 (CAPE [24]). These results highlight the strong generalization ability of GeneMAN to challenging poses.



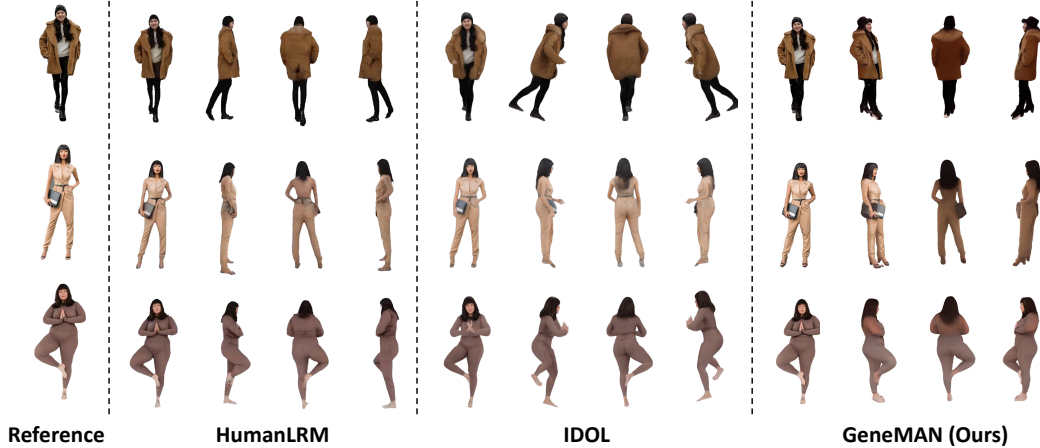


Figure 3: **Additional Comparison on in-the-wild Cases with HumanLRM [44] and IDOL [53].** We conduct experiments on the examples reported by HumanLRM. Best view with zoomed in.

## 137 D.2 Quantitative Geometric Evaluation

138 Following prior work, we perform a quantitative geometric comparison with baseline methods [33,  
 139 52, 46, 45, 50, 12, 8, 51] on the CAPE [24] dataset to assess geometry reconstruction quality.  
 140 Specifically, we report two commonly used metrics: Chamfer Distance (CD) and Point-to-Surface  
 141 distance (P2S), both measured in centimeters, between the ground truth scans and the reconstructed  
 142 meshes. Additionally, to assess the fidelity of reconstructed local details, we calculate the  $\mathcal{L}_2$  error  
 143 between normal images rendered from the reconstructed and GT surfaces, referred to as Normal  
 144 Consistency (NC). These renderings are obtained by rotating the camera around the surfaces at angles  
 145 of  $\{0^\circ, 120^\circ, 240^\circ\}$  relative to the frontal view.

146 However, it is crucial to highlight that previous template-based methods [52, 46, 45, 50, 12, 8, 51]  
 147 directly utilize ground truth SMPL [21, 27] provided in CAPE for evaluation. However, estimating  
 148 body shape and pose parameters from a single image is an ill-posed problem due to ambiguity,  
 149 leading to multiple possible solutions, as illustrated in Fig. 4. This presents an unfair advantage over  
 150 our template-free approach. Moreover, our work focuses on reconstructing 3D humans with high  
 151 fidelity from in-the-wild images, where accurate SMPL estimates are unavailable. To provide a fair  
 152 comparison, we re-evaluate the baselines on the CAPE dataset under an inference mode, where *GT*  
 153 *body poses are not provided as input*. The lack of access to such ground truth estimations leads to a  
 154 notable performance drop for template-based methods, which deviates from the results presented in  
 155 the original papers. The quantitative results are summarized in Tab. 2, where our method outperforms  
 156 the compared approaches across all geometric metrics, demonstrating the superior reconstruction  
 157 quality of our method.

Table 2: **Geometric Comparison with State-of-the-art Methods on the CAPE [24] dataset.** The best results are highlighted in **bold**. The second-place results are underlined.

Methods	CD ↓	P2S ↓	NC ↓
PIFu [33]	2.5580	2.5770	0.0925
PaMIR [52]	2.5502	2.5920	0.0925
ICON [46]	2.4147	2.4581	0.0872
ECON [45]	2.0782	<u>2.0296</u>	0.0798
GTA [50]	2.6785	<u>2.7760</u>	0.0914
TeCH [12]	2.3217	2.4163	0.0935
SiTH [8]	<u>1.9182</u>	2.0427	<u>0.0726</u>
SIFU [51]	2.5226	2.5692	0.0875
GeneMAN (Ours)	<b>1.8862</b>	<b>1.9724</b>	<b>0.0712</b>

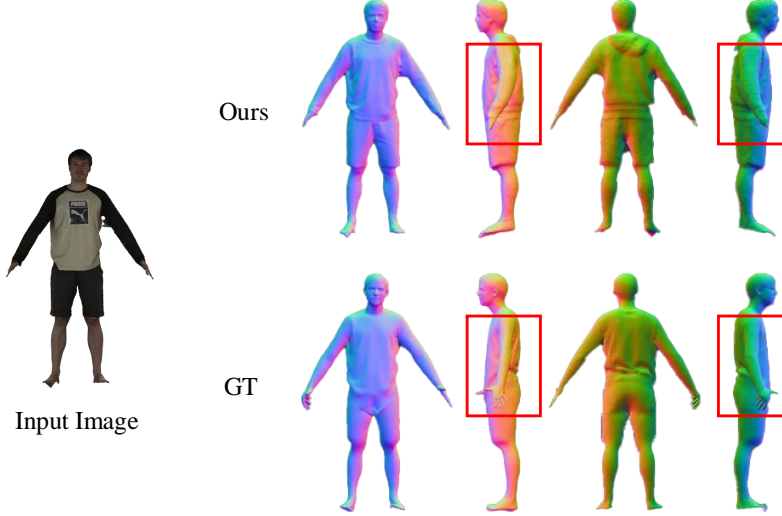


Figure 4: **Visualization Results of the Reconstructed Geometry on the CAPE [24] dataset.** While our template-free method successfully recovers plausible shapes, the orientation of the arms does not perfectly align with the ground truth, as highlighted by the red bounding box. This misalignment is difficult to avoid due to the inherent ambiguity, which can result in large 3D errors. Consequently, it is unfair to compare our approach with SMPL-based methods that utilize GT body parameters directly.

### 158 D.3 Ablation of Multi-Source Human Dataset

159 To validate the effectiveness of our constructed multi-source human dataset, we perform an ablation  
 160 study on each component: 3D scan human data, video human data, augmented human data, and 2D  
 161 human data. Specifically, we compare the reconstruction results using prior models trained with  
 162 various combinations of these data components. The following experimental settings are evaluated:  
 163 (a) “Baseline”, which replaces 2D and 3D prior models in GeneMAN with their original counterparts  
 164 (Stable Diffusion V1.5 [31] and Zero-1-to-3 [19]); (b) “Baseline + 3D”, which uses prior models  
 165 trained solely on 3D scanned human data; (c) “Baseline + 3D + Video”, which employs prior models  
 166 trained on both 3D scanned human data and video human data; (d) “Baseline + 3D + Video +  
 167 AUG” which incorporates prior models trained on 3D scan human data, video human data and  
 168 augmented human data; (e) “Ours”, which utilizes prior models trained on the complete dataset.  
 169 As detailed in Sec. 5.2 of the main text, we evaluate the performance of each setting using PSNR,  
 170 LPIPS, and CLIP-Similarity across a set of 30 test cases. As shown in Tab. 3, incorporating 3D  
 171 scans significantly improves the model’s multi-view consistency, resulting in a 0.046 increase in  
 172 CLIP-similarity. Furthermore, adding video data and augmented data enhances both texture quality  
 173 and consistency. By utilizing the full dataset, our full-fledged method achieves the best performance  
 174 in both texture quality and consistency.

Table 3: **The Effectiveness of Multi-source Human Dataset.** The best results are highlighted in bold.

Methods	PSNR $\uparrow$	LPIPS $\downarrow$	CLIP-Sim $\uparrow$
Baseline	31.503	0.018	0.662
Baseline + 3D	30.873	0.017	0.708
Baseline + 3D + Video	31.326	0.015	0.713
Baseline + 3D + Video + AUG	31.205	0.015	0.722
GeneMAN (Ours)	<b>32.238</b>	<b>0.013</b>	<b>0.730</b>

Table 4: **Model Efficiency.** Comparison of inference times between our model and the baselines. Note that PIFu [33], GTA [50], and SiTH [8] are feed-forward methods, whereas TeCH [12] and GeneMAN (ours) are optimization-based approaches that require per-subject optimization. Our GeneMAN model requires a total of 1.42 hours to generate a 3D human asset, with the following time breakdown: Geometry Initialization (22 min), Geometry Sculpting (15 min), Latent Space Texture Refinement (37 min), and Pixel Space Texture Refinement (11 min). All methods are tested on a single NVIDIA A100 80GB GPU.

	PIFu [33]	GTA [50]	TeCH [12]	SiTH [8]	GeneMAN(Ours)
<b>Inference Time</b>	4.6s	24s	3.5h	2min	1.42h

## 175 E Limitations

176 Although our method achieves superior reconstruction performance on in-the-wild images compared  
 177 to state-of-the-art approaches, it requires a longer optimization time to generate each 3D asset  
 178 compared to feed-forward methods such as PIFu [33], GTA [50], and SiTH [8], which may limit its  
 179 practical applicability. Nevertheless, our approach remains faster than TeCH [12], offering a 59.4%  
 180 increase in efficiency while delivering better quality. The model efficiency of both baseline methods  
 181 and ours are reported in Tab. 4. Regarding reconstruction quality, our method still struggles to achieve  
 182 fine-grained modeling for certain parts, such as the hands of full-body humans. Additionally, it  
 183 lacks specific designs for handling occluded individuals. For human-with-objects reconstruction,  
 184 our approach primarily focuses on reconstructing personal belongings, but it performs poorly with  
 185 particularly large or complex objects, such as bicycles. Research on human-with-object reconstruction  
 186 will be a key focus of our future work.

## 187 F Broader Impacts and Safeguards

188 GeneMAN introduces a generalizable framework for reconstructing high-quality 3D human models  
 189 from a single in-the-wild image. This method is capable of handling various body proportions,  
 190 poses, clothing, and personal belongings, offering a wide range of applications, including virtual  
 191 reality (VR), augmented reality (AR), telepresence, digital human interfaces, film production, and  
 192 3D game development. However, the widespread application of GeneMAN may also pose risks.  
 193 For instance, 3D human reconstruction technology could be misused to generate synthetic content,  
 194 which may raise ethical and legal concerns. To ensure that the application of GeneMAN is positive  
 195 and sustainable, we have implemented the following safeguards: 1) Ensure that the technological  
 196 outcomes of GeneMAN are equitably accessible to diverse social groups, including marginalized  
 197 communities. 2) Actively collaborate with communities and stakeholders to ensure that technological  
 198 applications meet societal needs and expectations. 3) Ensure that the development and application  
 199 of GeneMAN comply with all relevant laws and regulations, including data protection laws and  
 200 intellectual property rights. 4) Respect all intellectual property rights and ensure that all used data  
 201 and methods have been appropriately authorized.

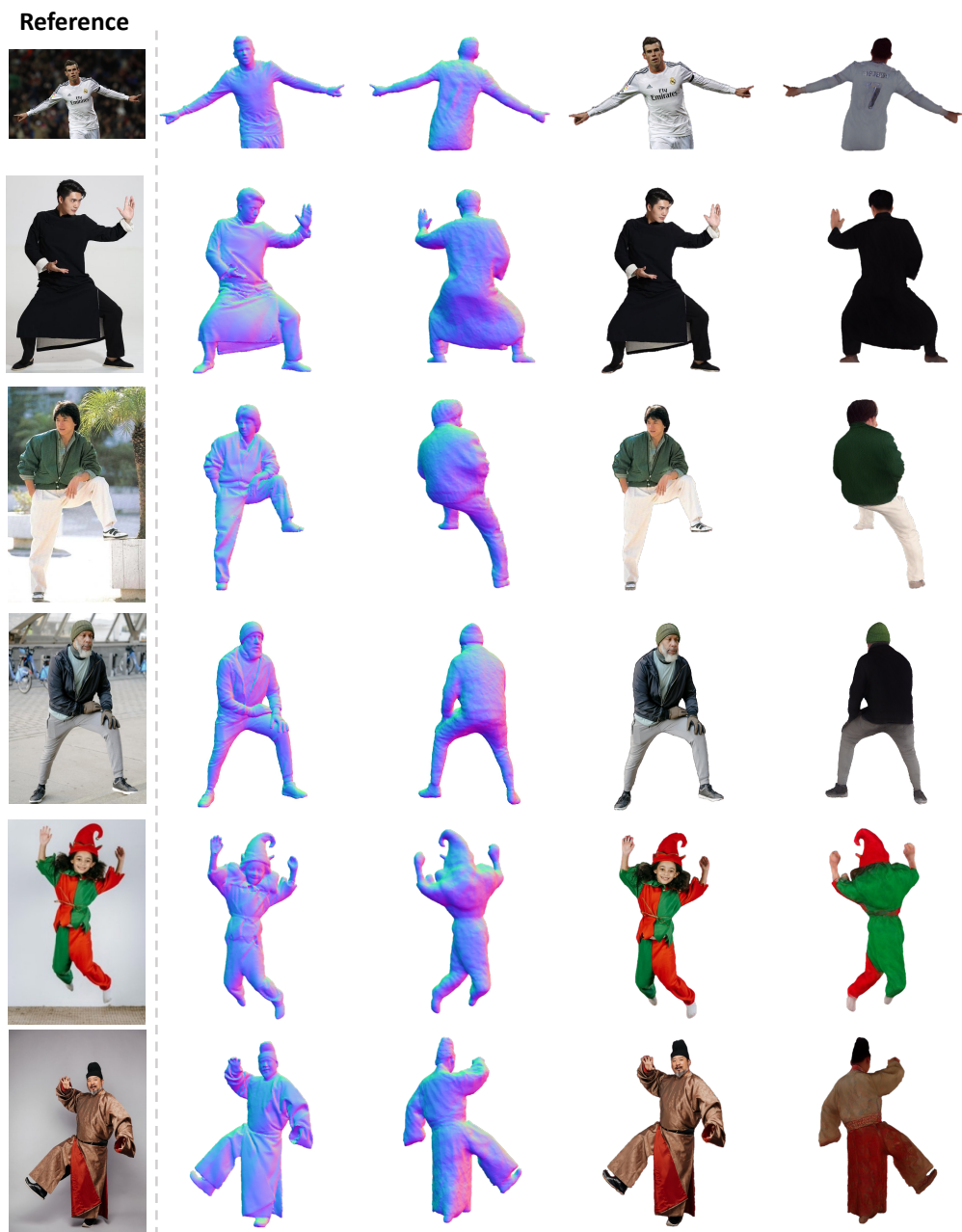


Figure 5: **More GeneMAN Results with Complex Poses on in-the-wild Images.** Best view with zoomed in.

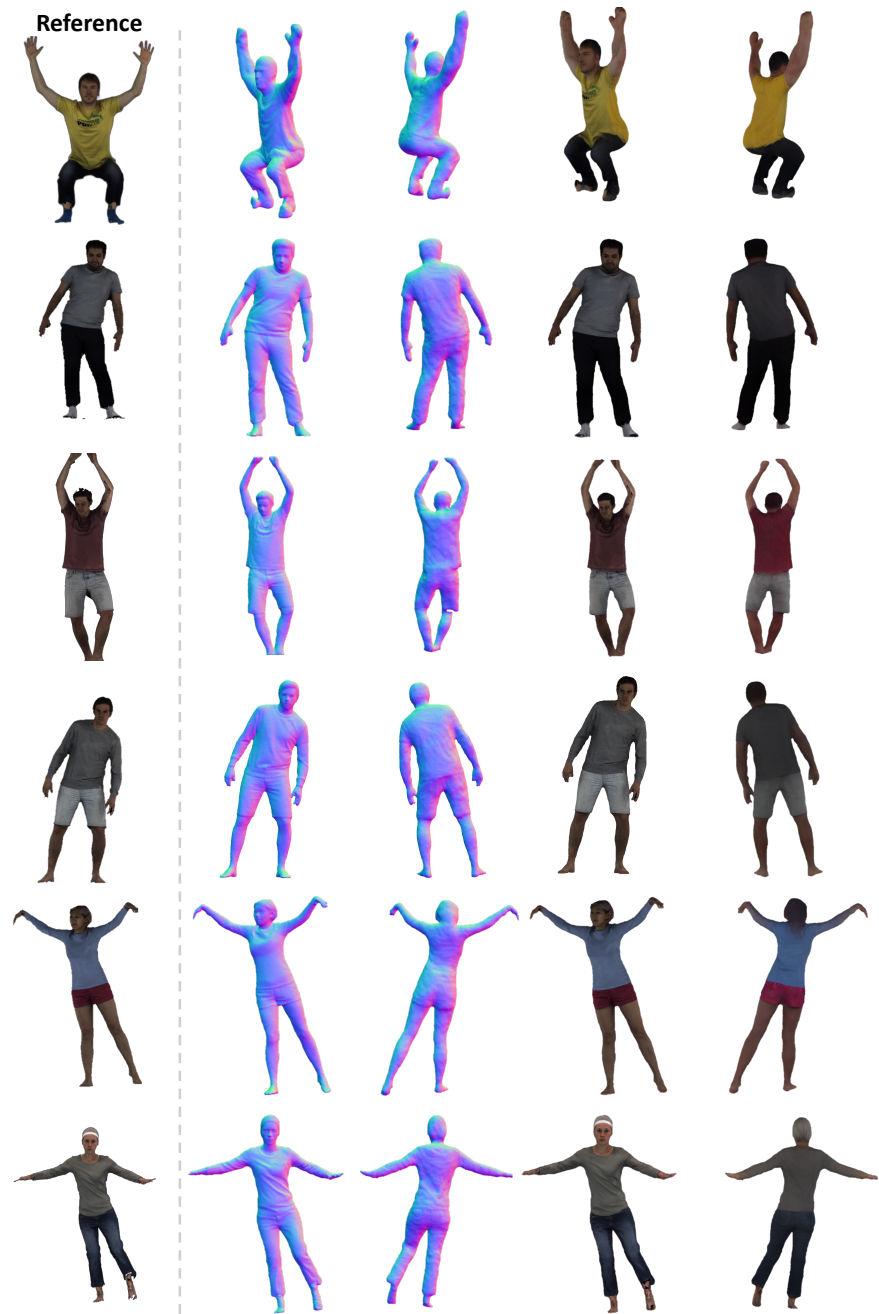


Figure 6: More GeneMAN Results with Complex Poses on CAPE [24]. Best view with zoomed in.



Figure 7: **Geometric Comparison on in-the-wild Images.** Best view with zoomed in.



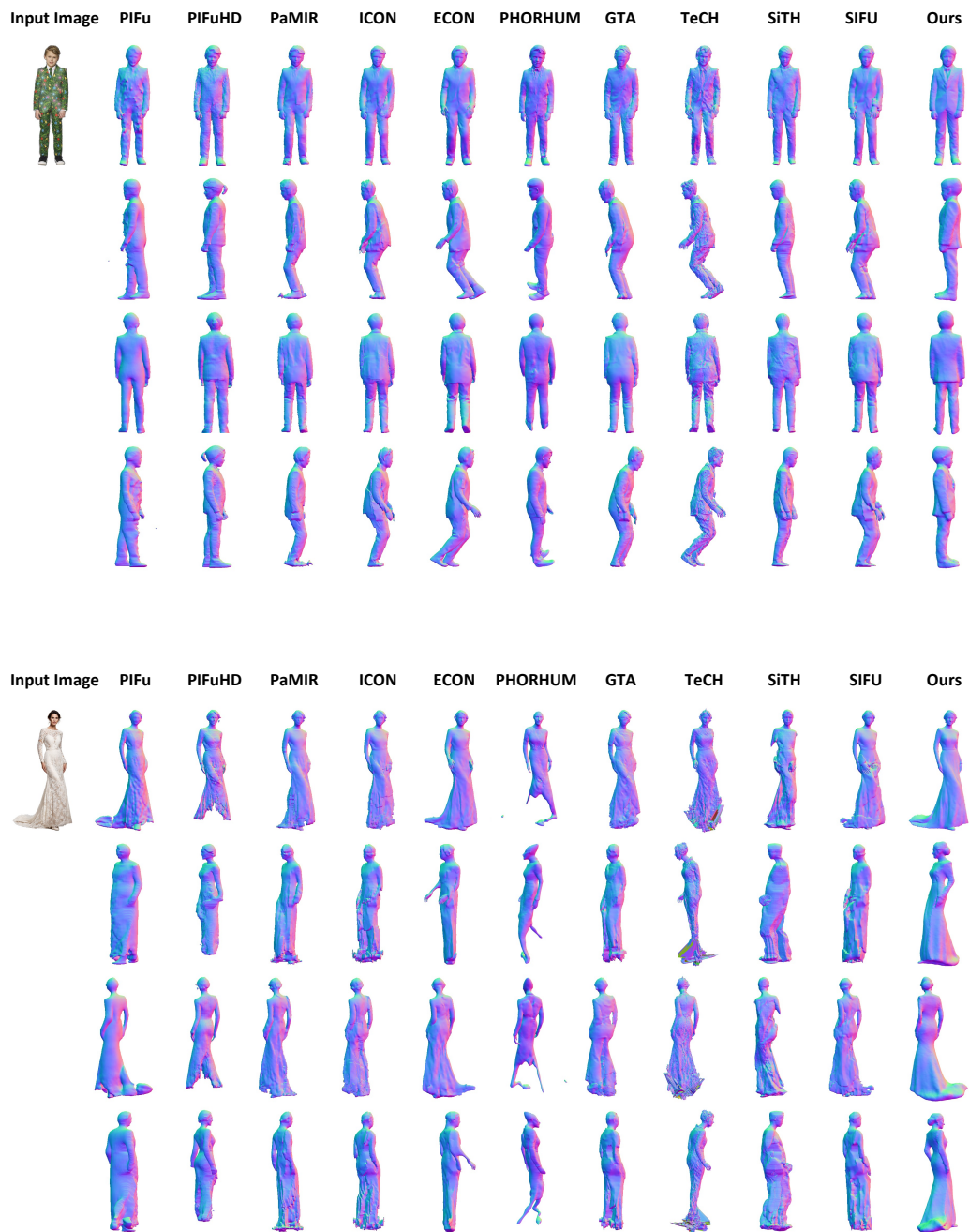


Figure 8: **Geometric Comparison on in-the-wild Images.** Best view with zoomed in.



Figure 9: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.





Figure 10: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.

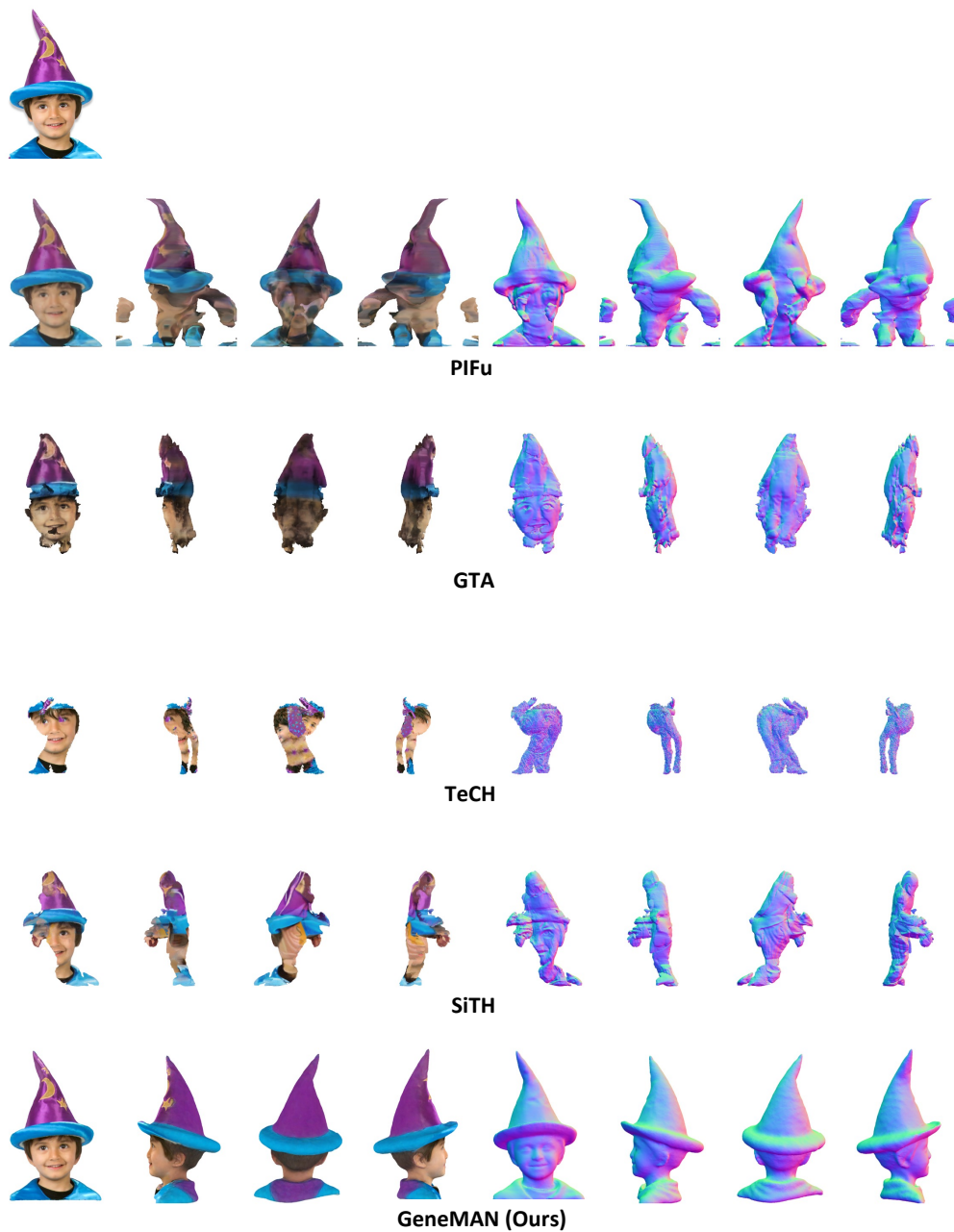


Figure 11: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.



Figure 12: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.



Figure 13: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.

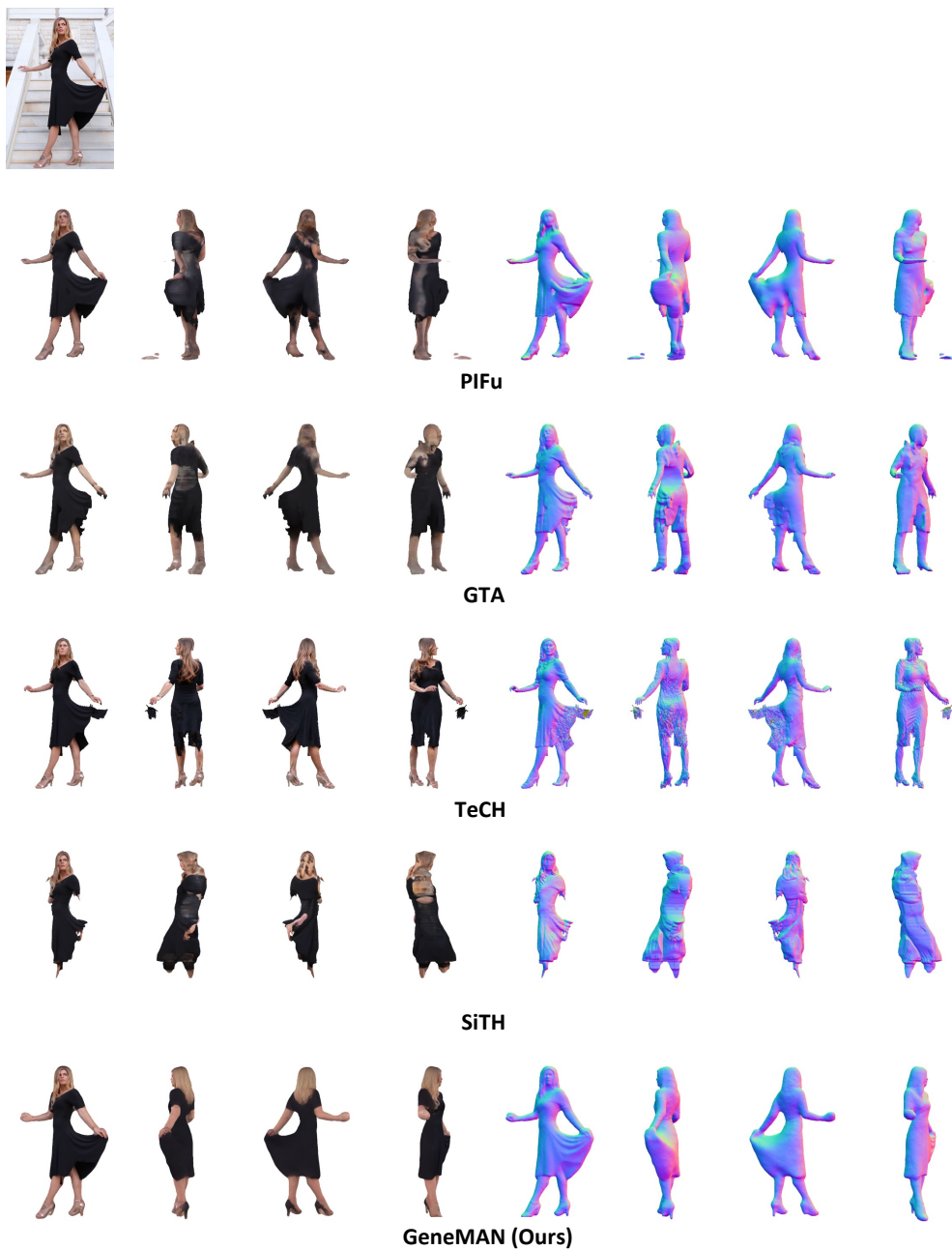


Figure 14: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.



Figure 15: **Qualitative Comparison on in-the-wild Image.** Best view with zoomed in.

## References

- [1] Renderpeople. <https://renderpeople.com>, 2018.
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2022.
- [4] 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*, pages 557–577. Springer, 2022.
- [5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19982–19993, 2023.
- [6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.
- [7] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21024–21035, 2023.
- [8] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 538–549, 2024.
- [9] Fangzhou Hong, Jiaxiang Tang, Ziang Cao, Min Shi, Tong Wu, Zhaoxi Chen, Tengfei Wang, Liang Pan, Dahua Lin, and Ziwei Liu. 3dtopia: Large text-to-3d generation model with hybrid diffusion priors. *arXiv preprint arXiv:2403.02234*, 2024.
- [10] 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. *arXiv preprint arXiv:2311.04400*, 2023.
- [11] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2024.
- [12] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024.
- [13] 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):1–12, 2023.
- [14] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.
- [15] Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6955–6965, 2024.
- [16] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.
- [17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

- [18] 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):1–16, 2021.
- [19] 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.
- [20] 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.
- [21] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.
- [22] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [23] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36, 2024.
- [24] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6469–6478, 2020.
- [25] 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*, pages 9981–9991, 2024.
- [26] 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):1–15, 2022.
- [27] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.
- [28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.
- [29] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [30] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023.
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [33] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019.
- [34] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 84–93, 2020.
- [35] 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*, 35:25278–25294, 2022.



- [36] Nikita Selin. CarveKit. [github.com/OPHoperHPO/image-background-remove-tool](https://github.com/OPHoperHPO/image-background-remove-tool), 2023.
- [37] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16911–16921, 2023.
- [38] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021.
- [39] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.
- [40] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [41] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1581–1593, 2022.
- [42] Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. Dreamcraft3d: Hierarchical 3d generation with bootstrapped diffusion prior. *arXiv preprint arXiv:2310.16818*, 2023.
- [43] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7464–7475, 2023.
- [44] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Single-view 3d human digitalization with large reconstruction models. *arXiv preprint arXiv:2401.12175*, 2024.
- [45] 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*, pages 512–523, 2023.
- [46] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13286–13296. IEEE, 2022.
- [47] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgb-d sensors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5746–5756, 2021.
- [48] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023.
- [49] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [50] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *Advances in Neural Information Processing Systems*, 36, 2024.
- [51] 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*, pages 9936–9947, 2024.
- [52] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3170–3184, 2021.
- [53] 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. *arXiv preprint arXiv:2412.14963*, 2024.