PrimDiffusion: Volumetric Primitives Diffusion for 3D Human Generation — *Supplementary Material*

Zhaoxi Chen[1] **Fangzhou Hong**[1] **Haiyi Mei**[2] **Guangcong Wang**[1]
Lei Yang[2] **Ziwei Liu**[1,✉]
[1]S-Lab, Nanyang Technological University [2]Sensetime Research
{zhaoxi001, fangzhou001, guangcong.wang, ziwei.liu}@ntu.edu.sg
{meihaiyi, yanglei}@sensetime.com

This is the supplementary material for PrimDiffusion: Volumetric Primitives Diffusion for 3D Human Generation. We introduce the implementation details of each component and training configurations in Sec. C. Additional discussions and experiment results are presented in Sec. B. Furthermore, we demonstrate our method within the supplementary video (Sec. A) for a more comprehensive analysis.

A Supplementary Video

We provide a video with more visual results of our work. In specific, it contains:

- An overview of PrimDiffusion.
- Visualizations of the denoising process during generation.
- Qualitative results of 360° novel view synthesis.
- Qualitative results of novel pose generalization.

B Additional Discussions and Results

B.1 Problem Definitions of Different Methods

The setting in our paper is learning from multi-view images, where the distribution of camera viewpoint is uniformly balanced. This setting is commonly shared among diffusion-based 3D generative models [11, 21]. Note that, the GAN-based works [3, 6, 13] used for comparisons are originally proposed to learn 3D representation from image collections, where multi-view data is unavailable. Therefore, we do not evaluate our method on datasets like DeepFashion [8], which breaks our assumption. Nevertheless, all baseline methods are retrained on our dataset for fair comparisons. And we make careful adaptations to our problem setting for each method. For example, we remove the pose-guided sampling training strategy in EVA3D [6] as our camera poses have uniform distributions. And we adjust the camera parameters sampling process in StyleSDF [13] to match the camera pose distribution in our dataset.

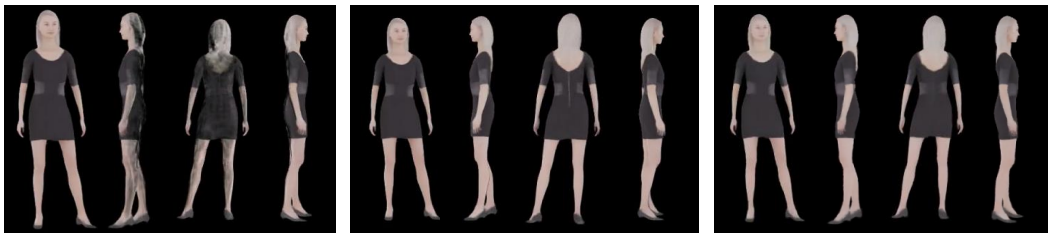
B.2 Robustness to SMPL Estimation

Previous 3D human generation methods [6, 24] require accurate SMPL [9] estimations to perform forward and inverse LBS from the observation space to the canonical space. In contrast, PrimDiffusion shows robustness to the error of SMPL estimation thanks to the independent degree of freedom for each primitive. Empirically, we perturb the local pose parameters of SMPL with noise sampled from $\mathcal{N}(0, 0.1)$, and retrain our method using these inaccurate SMPL parameters. The quantitative results are shown in Tab. 1. Although we train our method with inaccurate SMPL parameters, both the

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: **Robustness to SMPL [9] estimation.** The top three techniques are highlighted in red, orange, and yellow, respectively. †Methods trained on inaccurate SMPL estimations.

Methods	$\text{FID}_{\text{CLIP}} \downarrow$	FID $\downarrow$	KID$\times 10^2 \downarrow$	PCK $\uparrow$	Depth$\times 10^2 \downarrow$	w/ SMPL noise?
StyleSDF [13]	18.55	51.27	4.08 ± 0.13	-	49.37 ± 22.18	✗
EG3D [3]	19.54	24.32	1.96 ± 0.10	-	16.59 ± 21.03	✗
EVA3D [6]	15.03	44.37	2.68 ± 0.13	91.84	3.24 ± 9.93	✗
†Ours	12.34	23.65	2.16 ± 0.11	93.77	1.70 ± 1.42	✓
Ours	12.11	17.95	1.63 ± 0.09	97.62	1.42 ± 1.78	✗



w/ View Condition, w/ Decoder-free Rendering | w/ View Condition, w/o Decoder-free Rendering | **w/o View Condition, w/ Decoder-free Rendering**

Figure 1: **The impact of integrating view condition during generalizable primitive learning.** We separately train the encoder in the first stage with view conditions. We found that incorporating view conditions leads to view-specific primitives, resulting in novel view artifacts with decoder-free rendering (Left). These artifacts can be eliminated by repeatedly calling the forward pass of the model to output view-specific primitives for novel view synthesis (Middle). We solve this issue by removing the view condition, which enables reasonable novel view synthesis (indicating meaningful 3D representations) with decoder-free rendering (without model inference for novel views) (Right).

generation quality and geometry correctness surpass baseline methods even trained with accurate SMPL parameters. We attribute our robustness to the degree of freedom offered by independent kinematic parameters of primitives that correct the drift in SMPL estimation.

B.3 View Conditions

Previous 3D-aware generative models [3, 6] take as input the view condition or camera extrinsic to enable viewpoint control. However, we do not follow this practice. We argue that implicitly encoding view conditions will prevent models from: **1)** enabling decoder-free rendering, and **2)** learning meaningful 3D representations. It is obvious that encoding view conditions requires extra forward passes through the model for novel view synthesis, *i.e.,* the view condition from the novel viewpoint must feed into the network to render the corresponding view. Therefore, removing view conditions is critical for decoder-free rendering. Furthermore, we observe that view conditions can negatively impact 3D representation learning in our setting. To validate this argument, we implement a variant of our method by incorporating view conditions to the RGB mapping network F_{rgb} as its input. The visualizations are presented in Fig. 1. We found that incorporating view conditions leads to view-specific primitives, resulting in novel view artifacts with decoder-free rendering. These artifacts can be eliminated by repeatedly calling the forward pass of the model to output view-specific primitives for novel view synthesis. We solve this issue by removing the view condition, which enables reasonable novel view synthesis (*i.e.,* meaningful 3D representations) with decoder-free rendering (*i.e.,* without model inference for novel views).

B.4 Runtime Analysis

We report the computational cost of our model in this section. First, we provide the GPU memory consumption for both training and testing of the denoiser g_Φ in Tab. 2.

Moreover, we present the computational performance in light of the whole pipeline at inference time in Tab. 3. The FPS stands for novel view and pose synthesis of 300 frames. The amortized FPS indicates the FPS by considering the time of the denoising process. Our average inference FPS

Table 2: **GPU memory consumption.**

Phase	Mem (MB)	Batch size
Training	27428	4
Inference	21258	1

Table 3: **Amortized runtime analysis of the whole pipeline.** "A. FPS" indicates "Amortized FPS" which considers both the denoising and rendering time.

Inference mode	Denoising	FPS	A. FPS
DDIM, 100 steps	2.86 s	88.24	47.93
DDIM, 50 steps	1.45 s	88.24	61.86

Figure 2: **Qualitative results on THuman [26] dataset.** We use the same number of views as RenderPeople dataset to train our model on THuman dataset.

Figure 3: **Qualitative results of the baseline using triplane diffusion.** It tends to generate floating artifacts. We attribute the failure of triplane diffusion to the inefficiency of triplane in representing the human body which is a highly articulated object.

still outperforms baselines (the best baseline is only 22.97). More importantly, we only claim the real-time performance for novel view and novel pose synthesis once the denoising process is done. In most cases, the generative backbones account for identity-specific information. Since we disentangle the pose and view control from the generative backbone in a physically explicit way, we only need to call the denoiser once for identity-specific appearance. Thanks to our decoder-free rendering, we do not need any forward pass through the denoiser for novel view and novel pose synthesis, which is the fundamental reason for real-time rendering. However, existing 3D generative models like EG3D, EVA3D, and StyleSDF implicitly condition view and pose as input features, which forces them to call the forward pass of heavy generative backbones upon the view and pose changes.

B.5 More Qualitative Results

We provide additional qualitative results in Fig. 10. Please refer to the video for more comparison results. Note that, we observe that StyleSDF fails to explicitly control the viewpoints given the multi-view images training setting. And EG3D and EVA3D can explicitly control the viewpoints with 360 degrees. However, their renderings contain many artifacts.

Moreover, we also train our model on THuman 2.0 dataset [26]. Overall, we render 500 identities from Thuman 2.0 dataset with 36 camera views for each identity. We keep the training configuration unchanged and retrain our model from scratch on the rendered images. Note that, no explicit 3D supervision, e.g., normal or 3D mesh, is used during training. We present the results in Fig. 2 where the renderings show promising results.

In addition, we also perform a sanity check of the triplane as an alternative representation for the 3D human diffusion model. Specifically, we follow [21] to reconstruct triplane representation followed by a volumetric shared decoder. However, we found that it is not straightforward to fit triplanes and train the diffusion model on top of it for 3D human bodies. As shown in Fig. 3, triplane diffusion can easily generate floating artifacts around the human body. We attribute the failure of triplane diffusion on 3D humans to the inefficiency of triplane in representing the human body which is a highly articulated object. The human body only occupies a small portion of the space modeled by the triplane, thus the network wastes most of the parameters in modeling empty space. It poses challenges for both the fitting of triplanes and the convergence of diffusion models.

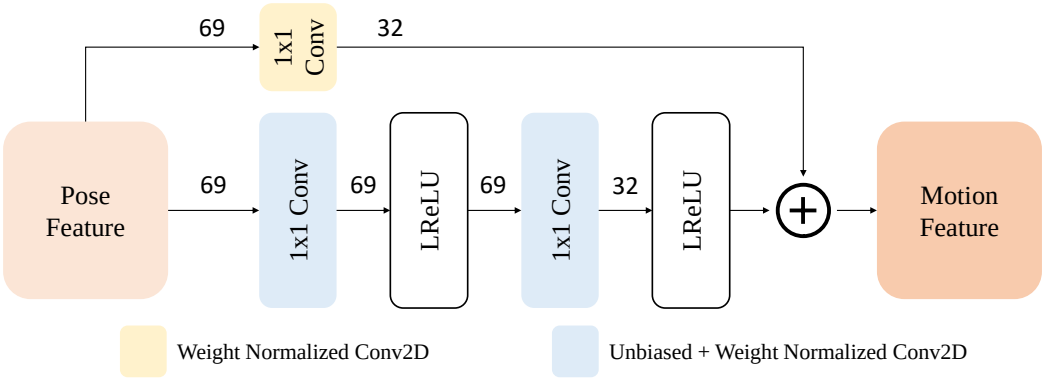


Figure 4: The network architecture of motion encoder F_θ. Both pose feature and motion feature 2D feature maps that preserve spatial information. The number on the arrow line denotes the number of channels. "LReLU" denotes LeakyReLU activation [23] with 0.2 negative slope.

C Implementation Details

C.1 Network Architecture

Encoder for Generalizable Primitive Learning. As introduced in Sec. 3.2 of the main paper, our proposed encoder for generalizable primitive learning consists of $\{F_I, F_\theta, F_{\text{rgb}}, F_\sigma, F_{\text{s}}\}$ with intermediate cross-modal attention layers. We document the implementation details of their architectures as follows:

- **Motion Encoder** F_θ. It takes as input the pose feature and outputs the corresponding motion feature. The architecture is illustrated in Fig. 4.
- **Image Encoder** F_I. It takes as input the UV-aligned image feature and outputs the corresponding appearance feature. The architecture is illustrated in Fig. 5.
- **Cross-Modal Attention Layer**. It takes as input the concatenated features from the motion branch and appearance branch and outputs the fused cross-modal feature. The architecture is illustrated in Fig. 6. Specifically, it consists of two basic attention blocks followed by a two-layer convolutional block.
- **RGB Mapping Network** F_{rgb}. It takes as input the fused cross-modal feature yielded by the attention layers, and outputs the color information $\mathbf{c}$ of primitives. The architecture is illustrated in Fig. 7.
- **Density Mapping Network** F_σ. It takes as input the fused cross-modal feature, and outputs the density information σ of primitives. The architecture is illustrated in Fig. 8.
- **Scale Mapping Network** F_{s}. It takes as input the fused cross-modal feature yielded by the attention layers, and outputs the delta scale factor $\delta\mathbf{s}$ of primitives. The architecture is illustrated in Fig. 9.

Denoiser g_Φ**.** We implement the denoiser as 2D U-Net with intermediate attention layers [16]. The model configuration is summarized in Tab. 5. Note that, the input shape corresponds to the volumetric primitive representation of one person $\mathcal{V}_0 \in \mathbb{R}^{[W\cdot S]\times[W\cdot S]\times[7\cdot S]}$ where we set $W = 32, S = 8$.

Table 4: **Statistics of clothing texture in the RenderPeople dataset.** The "Mono-color" denotes the texture with only one single color while "Complex texture" denotes the clothing with at least two different colors or patterns.

Body Part	Mono-color	Complex texture
Upper (shirts, jacket, etc.)	698	98
Lower (pants, trousers, etc.)	754	42

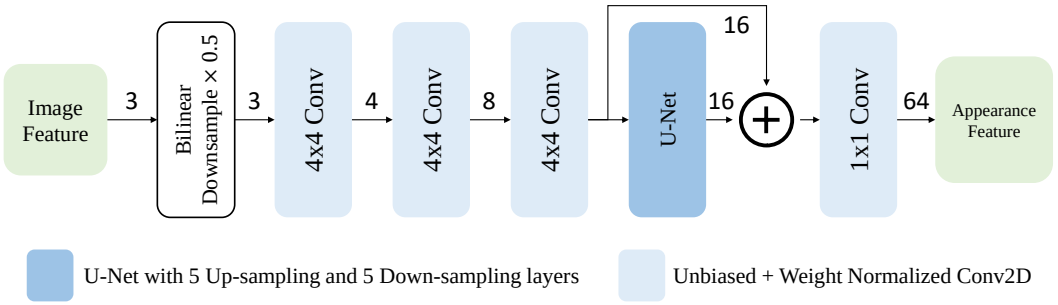


Figure 5: The network architecture of appearance encoder F_I. Both image feature and appearance feature 2D feature maps that preserve spatial information. The number on the arrow line denotes the number of channels. The "Bilinear Downsample" layer downsamples the input feature map by a factor of 0.5 via bilinear interpolation.

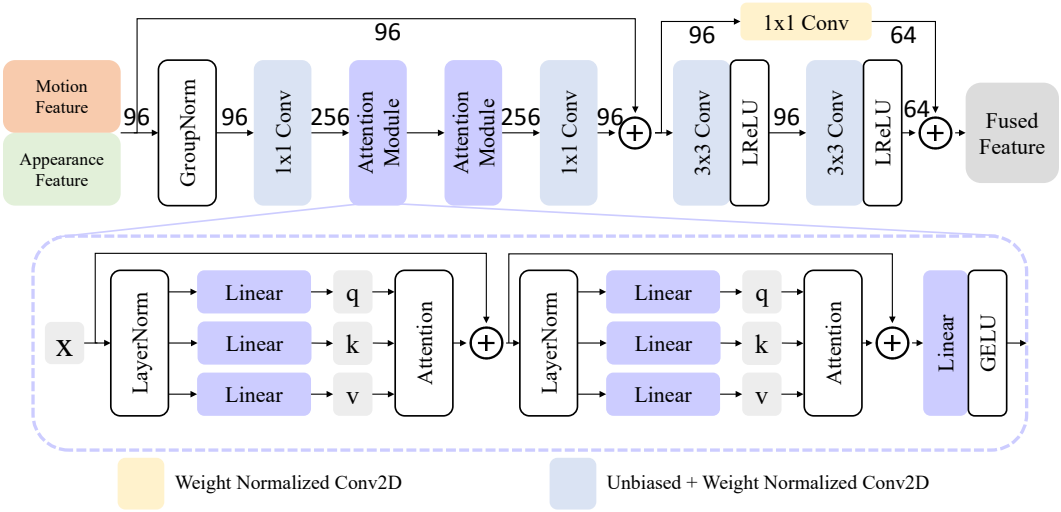


Figure 6: The architecture of cross-modal attention module. It takes as input the concatenated feature from the motion and appearance branch, and outputs the fused cross-modal feature map. The number on the arrow line denotes the number of channels. "GELU" denotes Gaussian Error Linear Units function [4]. "LReLU" denotes LeakyReLU activation [23] with 0.2 negative slope.

C.2 Dataset Details

We purchased 796 high-quality 3D humans from RenderPeople [20], where each sample is a 3D textured mesh obtained from high-resolution scans. Note that, we do not explicitly utilize the 3D supervision from the dataset, which is different from occupancy-based methods [17]. Instead, we render 36 camera views of each human subject, which are uniformly distributed on a circular trajectory around the human with a radius of 10 meters. The camera's orientation is set to point to the pelvis of the human body. Each rendered image has a 512×512 resolution with the corresponding field of view (FOV) of 14°. Furthermore, we utilize the motion retargeting technique to animate 3D human mesh, where each person is rendered with 20 different poses sampled from the AMASS [10] dataset. In total, the dataset used for training consists of $36 \times 20 \times 796 = 573120$ images.

Notably, the dataset has a strong bias towards mono-color clothes, which leads to the lack of complex texture in the generated results. We present the statistics of clothing texture in Tab. 4.

C.3 Training Hyperparameters

Generalizable Primitive Learning. We train the encoder $\{F_I, F_\theta, F_{\text{rgb}}, F_\sigma, F_{\text{s}}\}$ for generalizable primitive learning from multi-view images in an end-to-end manner. The learning objective is

Table 5: Model configuration of denoiser g_Φ. Note that, the input shape corresponds to the volumetric primitive representation of one person $\mathcal{V}_0 \in \mathbb{R}^{[W\cdot S]\times[W\cdot S]\times[7\cdot S]}$ where we set $W = 32, S = 8$.

Input shape	$256 \times 256 \times 56$
Input scaling factor	0.2
Diffusion steps	1000
Noise schedule	Linear, $\beta \in [1 \times 10^{-4}, 2 \times 10^{-2}]$
Channels	128
Depth	2
Channel multiplier	1, 2, 3, 4
Head channels	32
Number of attention heads	8
Transformer depth	1

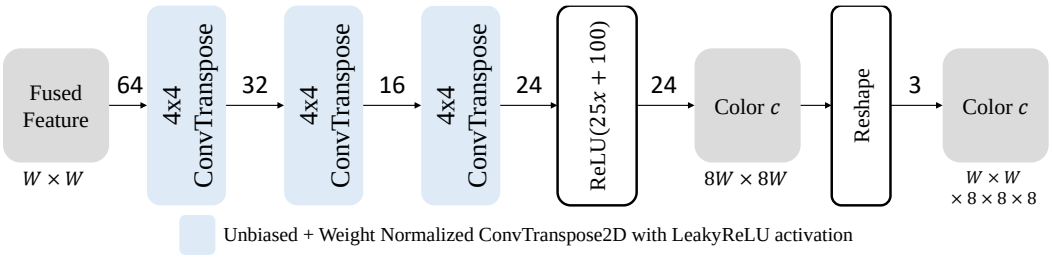


Figure 7: The architecture of mapping network F_{rgb}. It takes as input the fused cross-modal feature map and outputs the color information of primitives. The number on the arrow line denotes the number of channels.

presented as Eq. 3 in the main paper, where the loss weights are set as $\lambda_{\text{rgb}} = 1, \lambda_{\text{sil}} = 0.01, \lambda_{\text{vol}} = 0.001$, respectively. The silhouette loss $\mathcal{L}_{\text{vol}}$ is computed as L1-norm between the alpha mask and ground truth. We adopt Adam [7] optimizer with a learning rate of 5×10^{-5}. The training is distributed on 4 A100 GPUs with a batch size of 24 on each GPU for 300,000 iterations.

Primitive Diffusion. We train the denoiser g_Φ on 4 A100 GPUs with a batch size of 8 on each GPU for 200,000 iterations. The learning rate of Adam optimizer is set to 1×10^{-5}.

C.4 Evaluation Protocols

We introduce the evaluation protocols of metrics shown in Sec. 4 of the main paper, respectively. Unlike 2D image generation, 3D human generative tasks have three orthogonal dimensions to evaluate, *i.e.,* identities, poses, and views. Therefore, we uniformly randomly sample identity, pose, and view from the dataset to get ground truths for evaluation.

- **FID and KID.** Fréchet Inception Distance (FID) [5] and Kernel Inception Distance (KID) [2] are metrics for the quality of generated images. We utilize publicly available torch-fidelity[1] [12] to compute FID and KID against 48,000 images. The backbone model used to calculate the feature space distance is Inception-V3 [19]. All images are evaluated on a resolution of 512×512 with white backgrounds.
- **FID$_{\text{CLIP}}$.** In addition to FID, we utilize the image encoder of CLIP [14] to compute FID$_{\text{CLIP}}$. The backbone we leveraged is ViT-B/32[2].
- **PCK.** In order to evaluate the pose controllability of 3D human generative model, we evaluate the Percentage of Correct Keypoints (PCK) [1]. In specific, we use an off-the-shelf 2D human pose estimator [18] to predict 2D human poses from generated images. The ground truth poses are regressed from driven SMPL parameters and remapped to the format compatible with the pose estimator. The metric is computed against 5,000 images.

[1] https://github.com/toshas/torch-fidelity
[2] https://github.com/openai/CLIP/blob/main/model-card.md

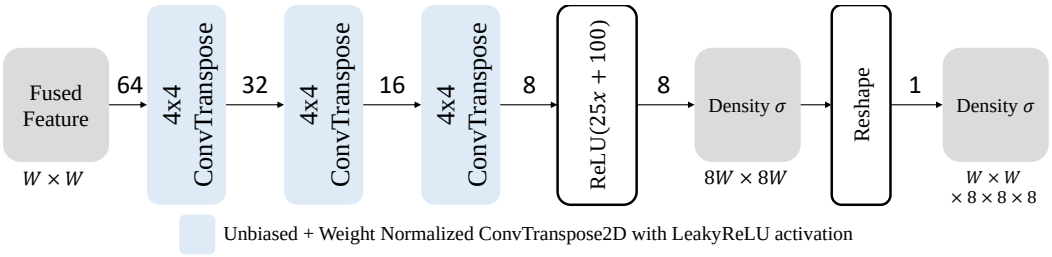


Figure 8: The architecture of mapping network F_σ. It takes as input the fused cross-modal feature map and outputs the density information of primitives. The number on the arrow line denotes the number of channels.

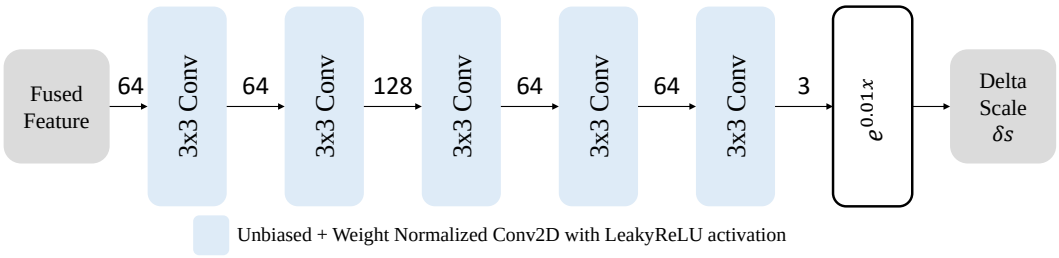


Figure 9: The architecture of mapping network F_s. It takes as input the fused cross-modal feature map and outputs the delta scale factor of primitives. The number on the arrow line denotes the number of channels.

- **Depth.** We follow a similar practice in EG3D [3] and EVA3D [6] for the evaluation of 3D geometry. We use a pre-trained model[3] [15] for monocular depth estimation to generate a pseudo ground truth depth map for each generated frame. The predicted depth map is generated via volume rendering by accumulating density for baseline methods. Finally, the depth error is computed as the L2 distance between the two. The metric is computed against 5,000 images.

Besides, the metrics for evaluating different design choices of volumetric primitives fitting (Tab. 3 of the main paper), *i.e.,* Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [22], and Learned Perceptual Image Patch Similarity (LPIPS) [25], are computed on rendered images in 512×512 resolution with black background masked out.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6

[3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 1, 2, 7

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 6

[3] https://github.com/isl-org/MiDaS

[6] Fangzhou Hong, Zhaoxi Chen, Yushi LAN, Liang Pan, and Ziwei Liu. EVA3D: Compositional 3D human generation from 2D image collections. In *International Conference on Learning Representations*, 2023. 1, 2, 7

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[8] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[9] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2

[10] 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*, pages 5442–5451, Oct. 2019. 5

[11] Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3D radiance field diffusion. In *arxiv*, 2022. 1

[12] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 6

[13] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, June 2022. 1, 2

[14] 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. 6

[15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 7

[16] 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. 4

[17] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019. 5

[18] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 6

[20] https://renderpeople.com/3d-people/. Renderpeople, 2018. 5

[21] Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3D digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4563–4573, 2023. 1, 3

[22] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[23] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 4, 5

[24] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3D generative model for animatable human avatars. In *Arxiv*, 2022. 1

[25] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[26] 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. 3

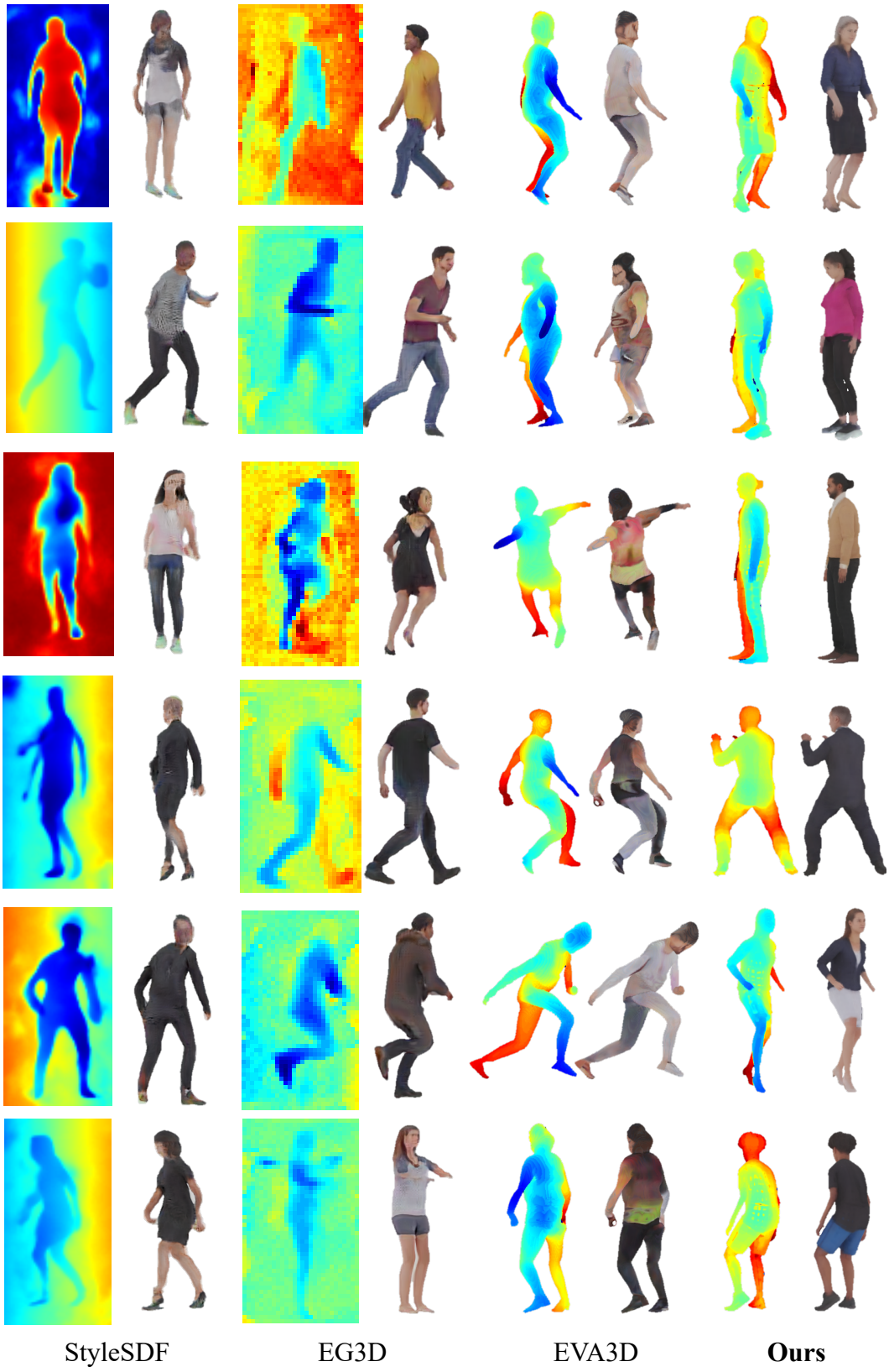


Figure 10: Additional qualitative results. Depth maps and RGB renderings are placed side-by-side.