Supplementary Materials for AG3D: Learning to Generate 3D Avatars from 2D Image Collections

Zijian Dong[1,2*] Xu Chen[1,3*] Jinlong Yang[3] Michael J. Black[3] Otmar Hilliges[1] Andreas Geiger[2]
[1]ETH Zürich, Department of Computer Science [2]University of Tübingen
[3]Max Planck Institute for Intelligent Systems, Tübingen

This **Supplementary Material** document provides additional details and experimental results as mentioned in the main paper. In addition, please see the **Supplementary Video**, which better illustrates the 3D results and also shows animation of the generated people.

1. Implementation

1.1. Technical Details

Canonical Generator: We use the same generator architecture as EG3D [2]. Similar to the camera pose conditioning in EG3D, we feed the sampled target human pose $\mathbf{p}$, represented as SMPL [9] joint angles, into the generator together with the latent code $\mathbf{z}$, in order to compensate pose-dependent non-linear deformations. To prevent overfitting, with a probability of 50%, this human pose conditioning is replaced with a random sampled pose from the training distribution.

Deformer: As described in the main paper, we use Fast-SNARF [3] to deform our canonical generation. The canonical skinning weights voxel grid values, which are used by Fast-SNARF to generate posed shapes, are derived from SMPL. The skinning weights of a grid point are derived by averaging the skinning weights of its 10 nearest canonical SMPL vertices, weighted by the inverse of the distance.

Volume Renderer: We render deformed 3D neural fields into images and 2D normal maps using volume rendering. Following NeRF [10], we adopt a two-pass importance sampling strategy to determine sampled point locations along the ray. We use 36 samples for the coarse stage and 36 samples for the fine stage. In addition, to improve rendering efficiency, we skip empty space based on the SMPL prior. To this end, we downsample SMPL vertices by a factor of 10 for computational efficiency and compute the distance of each sample point to its nearest neighbour in the downsampled vertex set. If the distance is above 0.15, we set the density of the sample to 0. We render images and normals at a resolution of 256^2.

Superresolution Module: To super-resolve our 256^2 renderings to 512^2 images, we first bilinearly upsample the renderings by $2\times$ and then feed the upsampled image into two convolution layers with a kernel size of 3.

Discriminators: Following EG3D [2], we apply the *image discriminator* at both resolutions: We upsample our low resolution rendering I, concatenate it with the super-resolved image I^+, and feed it to a StyleGANv2 [6] discriminator. For real images $\bar{I}$, we downsample and re-upsample them, and concatenate the results with the original image as input to the discriminator. Our *face image discriminator*, *normal discriminator*, and *face normal discriminator* are all standard StyleGANv2 discriminators, operating at resolutions of 64^2, 256^2 and 32^2 respectively. To guide the discrimination task, the pose parameters corresponding to the input images are given to all discriminators by modulating their convolution layers.

Training Objectives: Let D represent the set of discriminators and G denote the human generator.

$$\begin{aligned} V(D,G) = \lambda_{\text{eik}}\mathcal{L}_{\text{eik}} + \sum_{i\in T}\lambda_i[\mathbb{E}_{\mathbf{z}\sim\mathcal{Z},\mathbf{p}\sim\mathcal{P}} f(D_i(G(\mathbf{z},\mathbf{p}),\mathbf{p})) \\ +\mathbb{E}_{I_i\sim\mathcal{I}_i,\mathbf{p}\sim\mathcal{P}} f(-D_i(I_i,\mathbf{p})) - \lambda_{\text{R1}}\|\nabla D_i(I_i,\mathbf{p})\|^2] \end{aligned} \tag{1}$$

where $f(u) = -\log(1+\exp(-u))$, T is the set $\{\text{image}, \text{face\_image}, \text{normal}, \text{face\_normal}\}$ and $\mathcal{I}_{\text{image}}$, $\mathcal{I}_{\text{face}}$, $\mathcal{I}_{\text{normal}}$, and $\mathcal{I}_{\text{face\_normal}}$ are the distributions of the real data. $\mathcal{P}$ is the estimated pose distribution and $\mathcal{Z}$ is a standard Gaussian distribution.

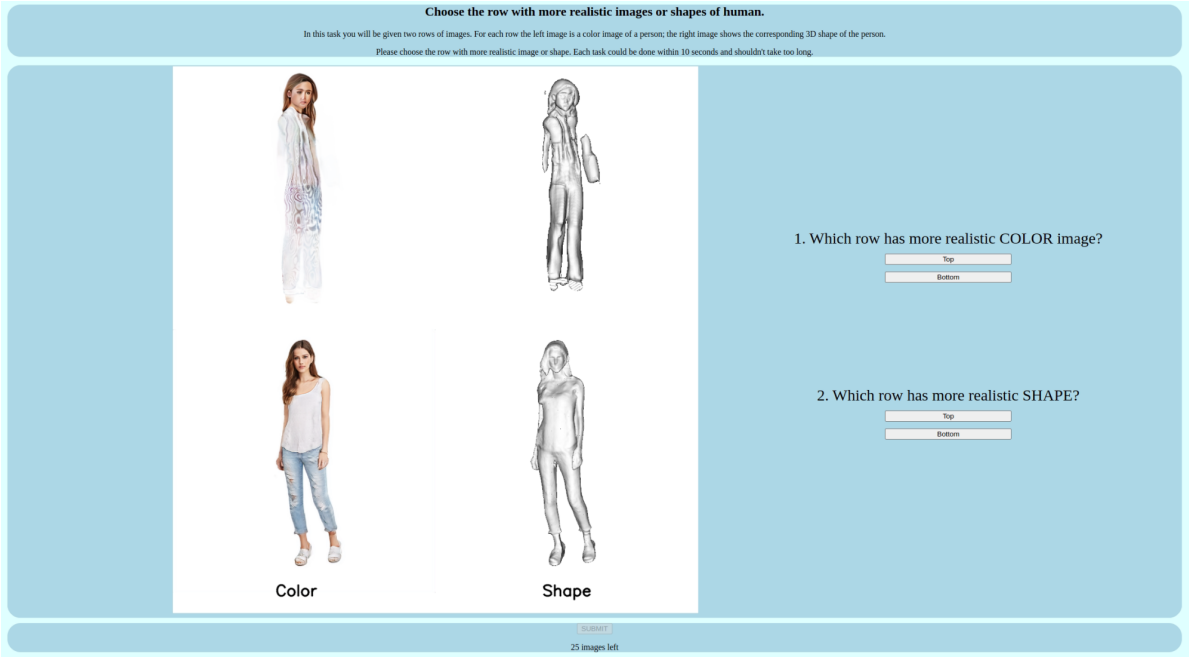


Figure 1: **The GUI for perceptual study with an exemplar comparison sample.**

We train our model in two stages. In the first stage, we set the loss weights to $\lambda_{\text{eik}} = 0.1, \lambda_{\text{R1}} = 5, \lambda_{\text{image}} = 1, \lambda_{\text{face\_image}} = 1, \lambda_{\text{normal}} = 0, \lambda_{\text{face\_normal}} = 0$. In the second stage, we add the normal discriminator and set $\lambda_{\text{normal}} = 0.05, \lambda_{\text{face\_normal}} = 0.05$.

1.2. Training Data Processing

For *DeepFashion* [8], we directly use the pre-processed subset with 8k images from [5] as our training data. *UBCFashion* [16] contains 500 sequences of fashion videos with subjects wearing loose clothing such as skirts. Following EVA3D [5], we treat these videos as individual images, ignoring temporal information. We extract 40K images from videos and use an off-the-shelf segmentation model [7] to remove backgrounds. Finally, the images are cropped and aligned according to the estimated human keypoints and then resized to 512^2. For both datasets, the corresponding SMPL parameters [9] are estimated by an off-the-shelf 3D pose estimator [17]. The 2D normal map of each image is obtained via a pre-trained human normal estimator [14].

2. Evaluation Details

2.1. FID computation

FID is used for evaluating the diversity and quality of human generation. Following EG3D [2], we compute FID for full images, and also for normal maps and face images. For each method to be evaluated, we generate 50000 samples in random human body poses and camera poses from the training set. The resolution of the generated images, normals and faces are 512^2, 256^2, 64^2 respectively. The pseudo-GT normal maps are obtained by running an off-the-shelf normal estimator [14] on the real images and downsampling the result to 256^2. The real face image is obtained by cropping the head regions of real images. We use an inception network [15] pre-trained on ImageNet [4] for all FID evaluations.

2.2. User Study

For practical applications, digital humans need to look like real people. To evaluate how faithfully our generated humans resemble real ones, we conduct a human perceptual study that compares the state-of-the-art method, EVA3D [5], and our method. The perceptual study is conducted via Amazon Mechanical Turk (AMT). We provide 50 participants with generated samples from our method and EVA3D. An example is shown in Fig. 1. In each comparison image, one row shows the shape and rendered color from our method and the other row shows those from EVA3D. The order of rows is randomly generated following a uniform distribution. For each comparison image, we ask the participant to choose: 1. the row that has a more realistic color image; and 2. the row that has a more realistic shape. In addition to 100 comparison samples that

each participant evaluates, we show the participants a tutorial image at the beginning, followed by 5 warm-up samples. We also include 10 catch-trial samples that are randomly placed among the 100 evaluation samples. Only the participants who pass more than 80% of the catch-trial samples are taken into final consideration. In total, we obtain approximately 4000 valid samples. Among these valid comparison evaluations, the participants prefer our shape 71.7% of the time, and prefer our color image 81.5% of the time.

3. Additional Qualitative Results

We show additional qualitative results to supplement qualitative result figures in the main paper.

3.1. Quality of 3D Human Generation

We first show additional results for novel view generation (Fig. 2), novel pose generation (Fig. 3) and interpolation (Fig. 4).

3.2. Comparison to Baselines

The additional comparison between other baselines and our method is shown in Fig. 5. The non-animatable models EG3D [2] and StyleSDF [12] generate worse human images and fail to learn reasonable geometry due to the difficulty of modeling highly variable articulated subjects. Since ENARF-GAN [11] can only be trained from low-resolution images, it struggles to generate high-resolution images. The current SotA method EVA3D [5] is the best among all of the baselines and is the focus of our comparison. Our method generally achieves better qualitative results with sharper images, more realistic geometry and more details. Our improvements are particularly pronounced when considering side views as shown in Fig. 6 and the modeling and deformation of loose clothing (Fig. 7 and Fig. 8).

3.3. Ablation study

We show additional qualitative results to supplement the analysis of normal discriminators and face discriminators in Fig. 9 and Fig. 10.

3.4. Text-guided clothing generation and editing

Our model is compatible with existing GAN control techniques [13]. We can achieve text-guided 3D human generation by optimizing the latent code to align the CLIP embedding of our output rendering and the embedding of the given text. We can further edit specific regions of the 3D human while keeping the remaining regions fixed by applying a local identity-preserving loss. Some results are shown in Fig. 11.

4. Discussions

4.1. Loose Clothing

Although neural field representations can effectively depict various clothing types including skirts, accurately deforming skirts, particularly in the area between the legs, remains a challenge for most articulation algorithms. Current methods typically deform neural fields of humans by obtaining a skinning weight field in *posed* space based on SMPL. The skinning weights of a point in *posed* space are computed by averaging the skinning weights of its K nearest SMPL vertices, weighted by the inverse of the distance. While this approach works well for points near SMPL surface, it struggles to accurately represent deformations of points far away from SMPL surface. As a result, when legs are widely separated, the areas in between, i.e. where the skirt is located, lack meaningful skinning weights. This leads to the splitting artifact in novel poses for methods that rely on SMPL nearest neighbours in posed space. We also derive skinning weights based on SMPL nearest neighbours, but do so in canonical space, where legs are not far apart. This produces smooth, meaningful and consistent skinning weights for skirts in canonical space. Our deformer, Fast-SNARF, then derives continuous deformations based on the canonical skinning weights, avoiding the splitting artifact. While our method better models skirt deformation, we note that the deformation is not physically correct. Future work could explore integrating physical simulation to generate realistic deformation of loose clothing.

4.2. Limitations and Future Work

Although achieving significant improvements, our method still has several limitations which are common to all existing articulation-aware GANs [1, 5, 11]. In the following, we discuss these limitations and potential solutions as future work.

Ambiguous Body Part Association: Our method as well as the SotA method EVA3D sometimes wrongly generate clothing patterns under the arm, or additional hands or arms on the torso, under unseen extreme poses (see Fig. 8a for EVA3D, Fig. 12a for ours). Since each training instance is observed only in one pose, the association of image observations to body parts cannot be uniquely determined when multiple parts occlude or contact each other (see Fig. 12b). Future work could investigate additional information to guide such associations, such as 2D dense correspondence predictions. Also, having multiple images of the same person in the same clothing (e.g. from video) would help the network avoiding this issue.

Rarely Observed Regions: Our method is capable of learning full 3D appearance and shape, as demonstrated by the results on the UBCFashion dataset (see also Supplementary Video). However, we found that both our method and EVA3D trained on the DeepFashion dataset produce artifacts for the back side, as shown in Fig. 13. This is because DeepFashion contains very few images of the back side, which makes it nearly impossible to learn plausible appearance and geometry without additional assumptions. To overcome this, an interesting direction is to incorporate multiple different data sources, such as multi-view images or 3D scans, which cover the complete appearance and shape, or to integrate symmetry priors.

Reliance on Trained Estimators: Existing articulation-aware GANs [1, 5, 11] including our method, rely on supervision provided by a 2D human segmentation model and a 3D SMPL estimator. Some inaccurate estimates from these models are likely to negatively impact our model. The ability to jointly refine these estimations during training, or to learn avatars without these estimators, might further improve the quality of the generated avatars.

Dataset Bias: Note that samples from generative models reflect the biases present in their training data. Existing 2D image collections used for training focus on fashion images from the internet which often lack diversity in skin tone, body shape, body pose, and age. Our work should be viewed as a methodological proof of concept. To avoid bias, deployed systems based on our approach should be trained on more diverse image data or data specific to a desired application.

Learning from in-the-wild images: While learning from ITW full-body images is indeed appealing, existing ITW datasets either contain limited subjects (e.g. 3DPW) or have complex backgrounds, varied lighting and clothing, extreme occlusions, and multiple people in each image. To date, no method including ours can handle all of these at once. Thus, existing methods focus on learning from fashion images, and even this has proved challenging. Our solution significantly improves the quality in this setting and is a meaningful step towards learning a more diverse model from large-scale ITW datasets or larger fashion datasets.

Learning from both 2D and 3D data: 3D data is useful, e.g. for learning pose-dependent wrinkles. However, such data is extremely scarce, thus not sufficient to capture the full diversity of humans and clothing alone. Ideally, one would combine both 3D and 2D data to learn a generative model with 3D data's quality and 2D data's diversity. Our goal is to push the limits of learning from 2D as this is foundational for future work that exploits both 3D and 2D data.

References

[1] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *arXiv preprint arXiv:2206.14314*, 2022. 3, 4

[2] 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*, pages 16123–16133, 2022. 1, 2, 3

[3] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *arXiv preprint arXiv:2211.15601*, 2022. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[5] 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. 2, 3, 4, 9, 10, 11

[6] 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. 1

[7] Yi Liu, Lutao Chu, Guowei Chen, Zewu Wu, Zeyu Chen, Baohua Lai, and Yuying Hao. Paddleseg: A high-efficient development toolkit for image segmentation, 2021. 2

[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 the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 2

[9] 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):1–16, 2015. 1, 2

[10] 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. *Communications of the ACM*, 65(1):99–106, 2021. 1

[11] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVII*, 2022. 3, 4

[12] 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*, pages 13503–13513, 2022. 3

[13] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[14] 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. 2

[15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[16] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. DWNet: dense warp-based network for pose-guided human video generation. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 51. BMVA Press, 2019. 2

[17] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

Figure 2: **Novel View Generation.** We generate 3D human appearance and shape, and render the resulting 3D representations from different viewpoints..

Figure 3: **Novel Pose Generation.** We generate 3D human appearance and shape, and render the resulting 3D representations in different body poses.

Figure 4: **Interpolation.** We show interpolated shapes and appearances between the leftmost and rightmost samples.

(a) Baselines

(b) EVA3D

(c) Ours

Figure 5: **Comparison with Baselines.** We show images and shapes generated by the baseline methods and our method. The results in (a) are taken from the EVA3D [5] paper.

(a) EVA3D

(b) Ours

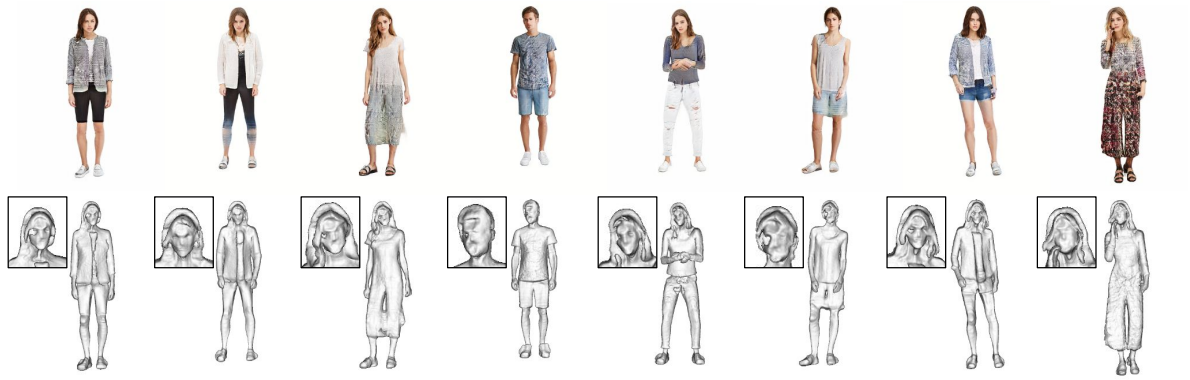
Figure 6: **Comparison with EVA3D on Novel Views.** We show frontal and side views of 3D humans generated by EVA3D [5] and our method.

(a) EVA3D (b) Ours

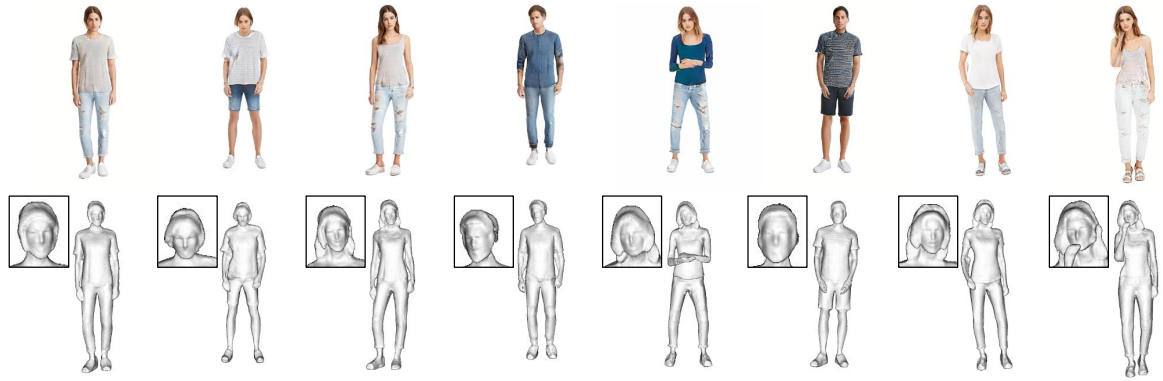
Figure 7: **Comparison with EVA3D on Loose Clothing.** We show 3D humans wearing skirts and dresses generated by EVA3D [5] and our method.

(a) EVA3D (b) Ours

Figure 8: **Loose Clothing Deformation.** Results with loose clothing in novel poses, generated by EVA3D and our method.



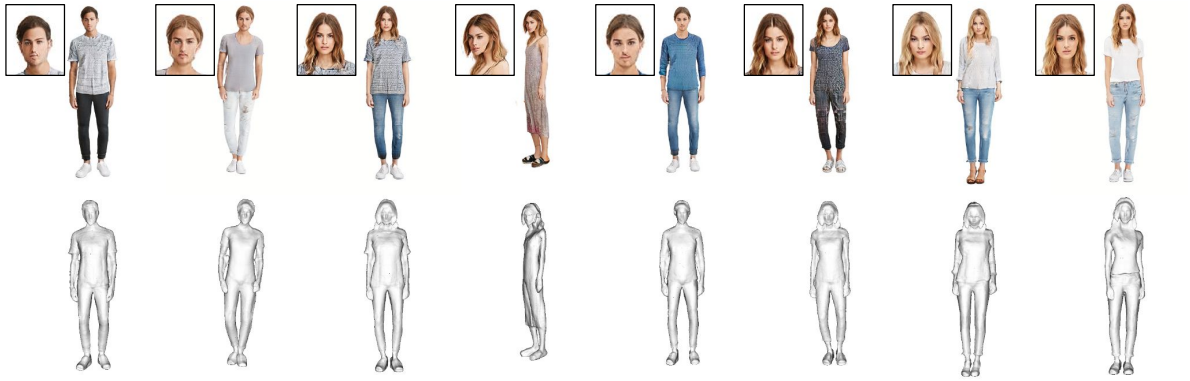
(a) Ours (w/o Normal GAN)



(b) Ours

Figure 9: **Ablation of the Normal Discriminator.** Results generated by our method with and without using the normal discriminator for training.

(a) Ours (w/o Face GAN)



(b) Ours

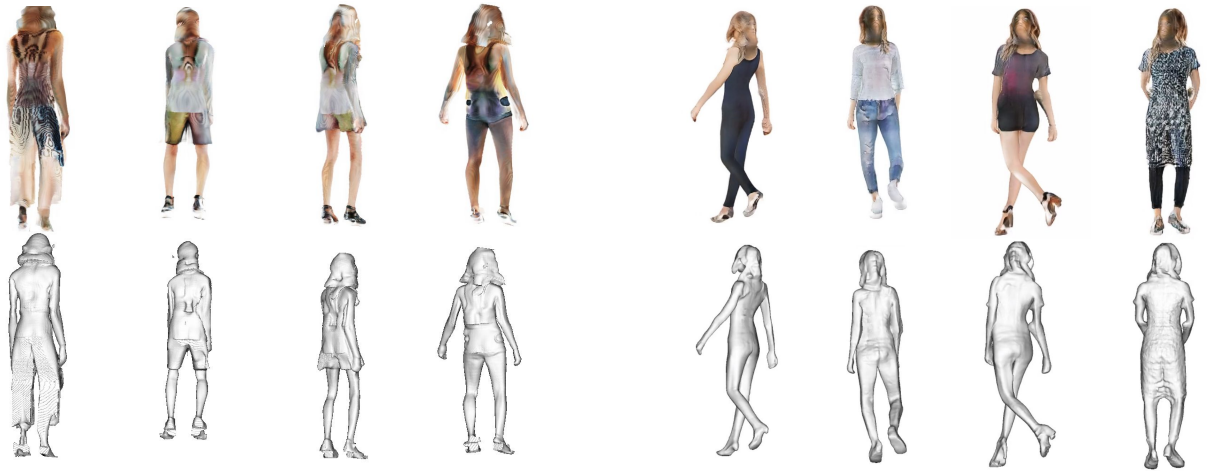
Figure 10: **Ablation of the Face Discriminator.** Results generated by our method with and without using the face discriminator for training.

Figure 11: **Text-guided Clothing Generation and Editing.** Results were generated by aligning the CLIP embedding of our output rendering and the embedding of the given text. The generation can be edited by applying a local identity-preserving loss.

(a) Ambiguous Body Part Association

(b) Potential Reason (low-resolution rendering)

Figure 12: **Ambiguous Body Part Association.** Our method sometimes generates clothing patterns under the arm, as shown in (a). These artifacts are often not visible in training poses, as shown in (b) left, due to close contact and occlusion, but are clearly visible in canonical space.



(a) EVA3D

(b) Ours

Figure 13: **Rarely Observed Regions.** When trained on the DeepFashion dataset, our method is better than EVA3D, but still suffers from artifacts on the back side (see head regions) due to the lack of back view training data.