EVA3D: COMPOSITIONAL 3D HUMAN GENERATION FROM 2D IMAGE COLLECTIONS — *Supplementary Material*

Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, Ziwei Liu ✉
S-Lab, Nanyang Technological University
{fangzhou001, zhaoxi001, yushi001, liang.pan, ziwei.liu}@ntu.edu.sg

This is the supplementary material for EVA3D. We provide more information of the training datasets in Sec. 1. More implementation details are introduced in Sec. 2. More visual results and comparisons are provided in Sec. 3. We also attached a demo video for better viewing experience.

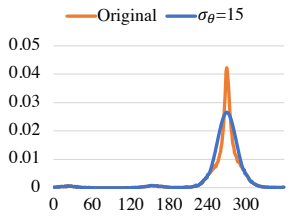
1 DATASETS

DeepFashion (Liu et al., 2016) collects fashion images from the internet. We only use images that contain the full body and not wearing dresses, which results in 8,036 images for training. We use SMPLify-X (Pavlakos et al., 2019) to estimate SMPL parameters and camera parameters. All images are resized to 512×256 for training. The alignment of the human body is the same as that proposed by Jiang et al. (2022).

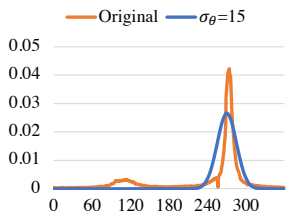
SHHQ (Fu et al., 2022) collects a larger-scale fashion dataset from the internet. It is processed similarly as DeepFashion, which results in 120,865 images in resolutions of 512×256. In our experiments, we find that models trained using SMPL estimated by SMPLify-X performs better than that of SPIN (Kolotouros et al., 2019). The head direction distribution of SHHQ, like DeepFashion, is also heavily imbalanced, as shown in Fig. 1 a). The accompanying blue line is the distribution of the proposed pose-guided sampling.

UBCFashion (Zablotskaia et al., 2019) is a fashion video dataset containing 500 sequences of models posing in front of the camera. Most models wear dresses in this dataset. We estimate SMPL sequences from videos by VIBE (Kocabas et al., 2020). We use all frames of the 500 videos and crop them to 512×256 for training, which leads to 192,179 samples. Although most models spin in front of the camera, the head direction of UBCFashion is still heavily imbalanced, as shown in Fig. 1 b).

AIST (Tsuchida et al., 2019) is a multi-view human dancing video dataset that provides rich poses and accurate SMPL estimations. We directly use the dataset processing scripts provided by ENARF-GAN (Noguchi et al., 2022) and get 72,000 samples. Each sample is resized to 256×256 for training.



a) SHHQ



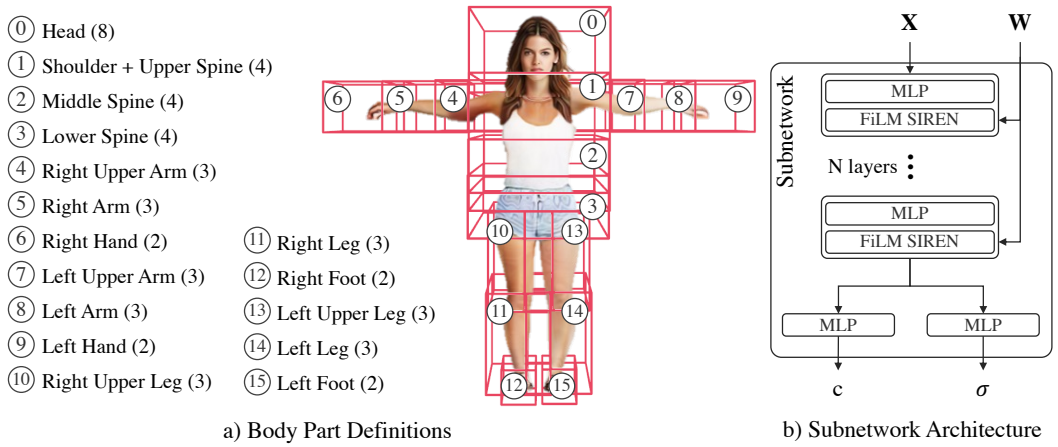
b) UBCFashion

Figure 1: Head Angle Distribution of SHHQ and UBCFashion.

2 IMPLEMENTATION DETAILS

2.1 NETWORK ARCHITECTURE

As introduced in the main paper, we split the whole body into 16 parts, which is shown in Fig. 2 in specific. For each part, a subnetwork is assigned, which is developed based on StyleSDF (Or-El et al., 2022). The architecture of each subnetwork is shown in Fig. 2 b). For each subnetwork, multiple MLP and FiLM SIREN (Chan et al., 2021) activation layers are stacked alternatively. At the end of each subnetwork, two branches are used to separately estimate SDF value and RGB value.



a) Body Part Definitions b) Subnetwork Architecture

Figure 2: a) shows our definition of 16 parts of human body. The number in the bracket is the number of corresponding subnetwork layers. b) shows the architecture of each subnetwork corresponding to each body part.

We assign different numbers of network layers for different body part empirically. Specific numbers are listed on Fig. 2 a). For the discriminator, we use the same architecture as that of StyleSDF (Or-El et al., 2022).

2.2 Training Settings

Hyperparameters. We use Adam optimizer (Kingma & Ba, 2014) for the optimization of both generator and discriminator. The learning rate for generator is 2×10^{-5}. The learning rate for discriminator is 2×10^{-4}. The loss weights as set empirically as $\lambda_{\text{off}} = 1.5$ and $\lambda_{\text{eik}} = 0.5$. For one ray, 28 query points are sampled. For the pose-guided sample, we choose to use $\sigma_\theta = 15°$.

R1 Scheduler. R1 regularization is used during training to penalize gradients of discriminator. Because it is highly challenging for the generator to learn plausible human appearance, the discriminator tends to overfit quickly if low R1 is set. But too high of R1 value would harm the final generation quality. Therefore, we set a R1 scheduler empirically, where R1 decrease from 300 to 18.5. R1 is cut in half every 50,000 iterations.

Augmentation. Inevitably, the SMPL estimations for 2D human images are not accurate for most samples. To compensate for the estimation error, we adopt small augmentations on real and fake samples before sent to the discriminator. The augmentation includes random panning, scaling and rotation in small ranges.

Runtime Analysis. The models are trained on 8 NVIDIA V100 GPUs for 5 days, with a batch size of 8. At test time, our model runs at ~ 5 FPS on one NVIDIA V100 GPU.

3 More Qualitative Results

Visual Comparison on UBCFashion & AIST. We further show renderings and corresponding meshes of three baseline methods and EVA3D trained on UBCFashion and AIST in Fig. 3. UBCFashion has dense views and simple human poses. Therefore, EG3D and StyleSDF succeeded in generating reasonable renderings. But the corresponding meshes lack details due to training at low native resolution (64×64). EVA3D gives the best visual results among the baseline methods and also generates plausible meshes with reasonable details. Due to complex human poses, StyleSDF fails on AIST. EG3D manages to generate reasonable 3D human, but fails to capture correct human structure in some cases. ENARF-GAN, for its low-resolution training, loses most details and generates rough meshes. EVA3D not only gets the best RGB renderings, but also generates meshes that preserve details like brims.

Qualitative Evaluations on Ablation Studies. As shown in Fig. 4, we visualize renderings and geometry generated by baseline methods described in the ablation studies in the main paper. The "Baseline", due to being trained at lower resolution (256×128), generates blurry renderings. The geometry fails to capture correct human structure (see broken knees). The compositional 3D human representation ("+ Composite") facilitates high resolution training (512×256). But lack of human prior leads to low-quality geometry (see unreasonable "wrinkles" on the upper bodies). By introducing a 3D human template and predicting delta SDF ("+ Delta SDF"), the visual quality increases and the geometry is mostly reasonable. However, the facial area is still flat due to the highly imbalanced viewing angle distribution. By using the pose-guided sampling ("+ Pose-Guided Sample"), we alleviate the imbalance issue and generate both high fidelity renderings and plausible geometry. To further validate our choice of Gaussian distribution in the pose-guided sampling, we visualize the results of models trained using uniform distribution ("+ Uniform Sample"). The middle of generated heads have severe artifacts.

More Qualitative Results of EVA3D. More qualitative results of EVA3D on four datasets are shown in Fig. 6, 7, 8, 9. For each sample, we show its novel view renderings and novel pose rendering.

4 More Discussions

4.1 More Results and Comparison of Interpolation and Inversion

We show more results of applications of EVA3D, which are latent space interpolation and inversion in Fig. 5. We also show latent space interpolation and inversion results on StyleSDF (Or-El et al., 2022) for comparison. Because StyleSDF has no control over human poses, pose changes are observed (see left hands) when interpolating between latent codes. The texture stitching problem (Karras et al., 2021) can also be seen during pose changing. Moreover, some middle results are semantically incorrect due to inferior generation quality of StyleSDF. In contrast, EVA3D demonstrate consistently high quality interpolation results during latent space interpolation. For the inversion, StyleSDF fails to recover detailed appearances of input target even with the second stage fine-tuning of PTI (Roich et al., 2021). Though the geometry are hard to be inverted with a single image as input, EVA3D manages to recover fine details of the input image. Admittedly, the inverted results are blurry, especially when changing camera views. We think that is because the ambiguity of a single image as input and errors introduced by single image SMPL estimation (Pavlakos et al., 2019). Besides, PTI (Roich et al., 2021), as an inversion method designed for 2D GANs, might not be suitable for 3D-aware GAN inversion. We expect more explorations (Cai et al., 2022; Lin et al., 2022) in 3D-aware GAN inversion in the future, which might inspire tasks like single image 3D reconstruction and 3D-aware editing.

4.2 Neck Line Artifact Issue on DeepFashion

As shown in Fig. 6, there are apparent geometric line artifacts around the neck areas. After careful comparison, we find that these line artifacts can only be found on the samples generated from DeepFashion training and can only be found on the necks and shoulders. Moreover, as shown in Fig. 4, the "Baseline" results do not have line artifacts. Based on the observations, we think the problem should be caused by DeepFashion, the compositional representation and head area modeling. Firstly, DeepFashion only contains $8,036$ samples, which is an order of magnitude less than other three datasets. It is known that generation models trained with small data are prone to artifacts (Karras et al., 2020). Moreover, unlike other datasets, many models in DeepFashion have long hair hanging in front of the shoulder, causing trouble for the compositional representation to model. The "head" and "shoulder" sub-modules try their best to composite the hanging-down hair and manage to give relatively continuous renderings. But due to the lack of 3D supervision, it is difficult for both modules to composite continuous surfaces at edges of their bounding boxes. The bigger limitation behind this observation is that the current compositional representation cannot model loose garments, accessories or other body parts (like hair). Using separate modules to handle loose parts (like hair) might alleviate the problem. Since it is not a trivial operation and out of the focus of this work, we will leave that to the future work.

Figure 3: Visual Comparison on UBCFashion & AIST. Zoom in for the best view.

Figure 4: Qualitative Evaluations on Ablation Studies. Zoom in for the best view.

Figure 5: **More Results of EVA3D Applications and Comparison with Baseline Methods.** a) Interpolation on the latent space gives smooth transition between two samples. b) Inversion results (right) of the target image (left).

Figure 6: More Qualitative Results of EVA3D on DeepFashion. Zoom in for the best view.

Figure 7: More Qualitative Results of EVA3D on SHHQ. Zoom in for the best view.

Figure 8: More Qualitative Results of EVA3D on UBCFashion. Zoom in for the best view.

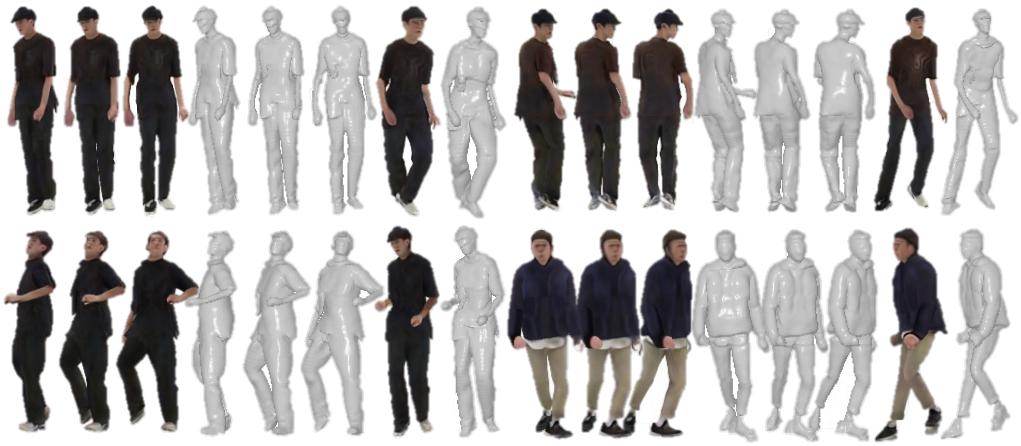


Figure 9: More Qualitative Results of EVA3D on AIST. Zoom in for the best view.

REFERENCES

Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional p-gan for single image to neural radiance fields translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3981–3990, 2022.

Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809, 2021.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint arXiv:2204.11823*, 2022.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. doi: 10.1145/3528223.3530104.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

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

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5253–5263, 2020.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2252–2261, 2019.

C.Z. Lin, D.B. Lindell, E.R. Chan, and G. Wetzstein. 3d gan inversion for controllable portrait image animation. In *ECCV Workshop on Learning to Generate 3D Shapes and Scenes*, 2022.

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.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. *arXiv preprint arXiv:2204.08839*, 2022.

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*, pp. 13503–13513, 2022.

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*, pp. 10975–10985, 2019.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pp. 501–510, Delft, Netherlands, November 2019.

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