StructLDM: Structured Latent Diffusion for 3D Human Generation (Supplementary Material)

Tao Hu, Fangzhou Hong, and Ziwei Liu

S-Lab, Nanyang Technological University, Singapore

1 Implementation

1.1 Network Architecture

Structured Auto-decoder. We adopt the decoder architecture of StyleSDF and EVA3D for each structured local NeRFs. For each subnetwork, multiple MLP, and FiLM SIREN activation [18] layers are stacked alternatively, and at the end of each subnetwork, two branches are used to separately estimate SDF value and RGB value. Different numbers of network layers for different body parts are assigned empirically: 4 layers for Head; 3 layers for Shoulder + Upper Spine, Middle Spine, Lower Spine; 2 layers for Right Upper Arm, Right Arm, Right Hand, Left Upper Arm, Left Arm, Left Hand, Right Upper Leg, Right Leg, Right Foot, Left Upper Leg, Left Leg, Left Foot with a similar design as EVA3D, whereas we adopt a more light-weight architecture with fewer layers for each subnetwork. The detailed diagram can be found in [6]. We render human features at 128×128 resolution by volume rendering using the structured auto-decoder.
Global Style Mixer. We utilize a receptive field of 4 in the experiments, *i.e.* upsampling the 128×128 renderings to 512×512 images. We employ two convolution blocks each containing a bilinearly upsampling step and two convolutional layers with a kernel size of 3 to upsample the images by a factor of 4.
Latent Size. Depending on the scale of the training dataset, the latent size is $128 \times 128 \times 16$ for UBCFashion [24], $128 \times 128 \times 24$ for RenderPeople [20], $64 \times 64 \times 24$ for THUman2.0 [22], and $64 \times 64 \times 32$ for DeepFashion [11].
Discriminator. We adopt the discriminator architecture of PatchGAN [3,7] for adversarial training. Note that different from EG3D [1] that applies the image discriminator at both resolutions, we only supervise the final rendered images with adversarial training and supervise the volumetric features with reconstruction loss.
Diffusion Model. Our diffusion model is based on the UNet architecture of [12], with four ResNet [5] blocks and a base channel of 128.

1.2 Optimization of Auto-decoder

StructLDM is trained to optimize renderers G_1, G_2 and structured embeddings $\mathcal{Z}$. Given a ground truth image I_{gt}, we predict a target RGB image I^{+}_{RGB} with the following loss functions:

Pixel Loss. We enforce an ℓ_1 loss between the generated image and ground truth as $L_{pix} = \|I_{gt} - I^+_{RGB}\|_1$.
Perceptual Loss. Pixel loss is sensitive to image misalignment due to pose estimation errors, and we further use a perceptual loss [8] to measure the differences between the activations on different layers of the pre-trained VGG network [17] of the generated image I^+_{RGB} and ground truth image I_{gt},

$$L_{vgg} = \sum \frac{1}{N^j} \left\| g^j \left(I_{gt} \right) - g^j \left(I^+_{RGB} \right) \right\|_2, \tag{1}$$

where g^j is the activation and N^j the number of elements of the j-th layer in the pre-trained VGG network.
Adversarial Loss. We leverage a multi-scale discriminator D [21] as an adversarial loss L_{adv} to enforce the realism of rendering, especially for the cases where estimated human poses are not well aligned with the ground truth images.
Face Identity Loss. We use a pre-trained network to ensure that the renderers preserve the face identity on the cropped face of the generated and ground truth image,

$$L_{face} = \| N_{face} \left(I_{gt} \right) - N_{face} \left(I^+_{RGB} \right) \|_2, \tag{2}$$

where N_{face} is the pre-trained SphereFaceNet [10].
Volume Rendering Loss. We supervise the training of volume rendering at low resolution, which is applied on the first three channels of I_F, $L_{vol} = \|I_F[:3] - I^D_{gt}\|_2$. I^D_{gt} is the downsampled reference image.
Latent Regularization. To allow better learning of the latent diffusion model, we regularize the latent with L2 regularization and TV loss.
Geometry Regularization Loss. Following EVA3D [6], we predict the delta signed distance function (SDF) to the body template mesh. Therefore, we penalize the derivation of delta SDF predictions to zero $\mathcal{L}_{\text{eik}} = E_x[\|\nabla(\Delta d(x))\|^2_2]$ [4].

The networks were trained using the Adam optimizer [9]. It takes about 3.5 days to train an auto-decoder from about 80K images on UBCFashion on 4 NVIDIA V100 GPUs, and about 3 days to train a diffusion model (based on [12]) with 4 NVIDIA V100 GPUs.

1.3 Training Data Precessing

UBCFashion [24] contains 500 sequences of fashion videos, and we uniformly extract about 80K images from these videos as training data. We render 24 multi-view images for RenderPeople [20] and THUman2.0 [22], which yield about 190K images and 12K images separately. For DeepFashion [11], we directly use the pre-processed subset with 8K images from EVA3D [6] as our training data. We crop and resize the images to 512×512 for training.

1.4 Part-aware Diffusion

Benefiting from the semantic design of latent space, StructLDM supports local editing by part-aware diffusion in inference. Given a part mask $\boldsymbol{M} \in [0, 1]$ and

Fig. 1: Trade-off between reconstruction fidelity and generation quality on DeepFashion [11]. Row 1: reconstruction results by auto-decoder. Row 2: latent visualization by Principal Component Analysis (PCA). Row 3: generation results.

reference latent $\boldsymbol{y}_0$, we generate $\boldsymbol{x}_{t-1}$ from noised $\boldsymbol{x}_t$ based on the estimated value for $\boldsymbol{x}_0$, namely $\bar{\boldsymbol{x}}_0^{(t)}$, which is refined to get $\hat{\boldsymbol{x}}_0^{(t)}$:

$$\min_{\hat{\boldsymbol{x}}_0^{(t)}} \left\| \hat{\boldsymbol{x}}_0^{(t)} - \bar{\boldsymbol{x}}_0^{(t)} \right\|_2 + \lambda \left\| (\mathbf{1} - \boldsymbol{M}) \odot \left(\boldsymbol{y}_0 - \bar{\boldsymbol{x}}_0^{(t)} \right) \right\|_2 \tag{3}$$

where $\odot$ denotes the Hadamard product, $\lambda = 0.5$ is a hyper-parameter controlling the balance between the diffusion prior and the degradation constraint. It admits a closed-form solution:

$$\hat{\boldsymbol{x}}_0^{(t)} = \frac{\bar{\boldsymbol{x}}_0^{(t)} + \lambda(\mathbf{1} - \boldsymbol{M})^2 \odot \boldsymbol{y}_0}{\mathbf{1} + \lambda(\mathbf{1} - \boldsymbol{M})^2} \tag{4}$$

which can be optimized using SGD when a closed-form solution is not feasible. The part-aware latent diffusion is similar to image inpainting in [23]. All the mathematical operations are pixel-wise.

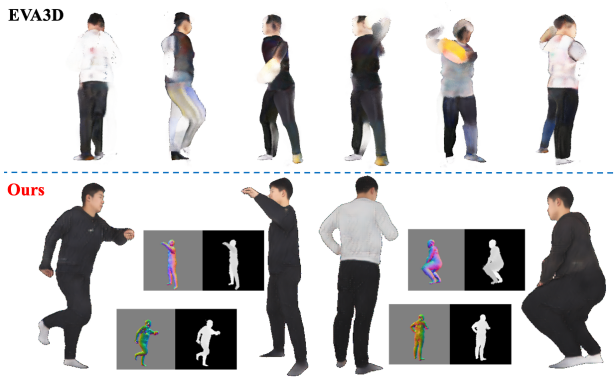

Fig. 2: Comparisons on THUman2.0 [22]. The geometry is visualized as normal/depth maps at 128×64 resolution.

Fig. 3: Qualitative generations on DeepFashion [11].

2 Additional Experimental Results

2.1 The Effect of Latent Mapping for Reconstruction and Generation

Autodecoding human latents from single images is challenging due to sparse observations. Fig. 1 shows the reconstruction results, latent visualizations, and generation results. It is observed that the learned latents are noisy and unfriendly for latent diffusion, i.e., the generations are noisy. Instead, we employ a mapping network to smooth the unobserved body parts in the latent space, which yields smoother latents and enables realistic generations. Though the latent mapping improves the generation quality, it degrades the reconstruction fidelity, which imposes challenges of learning auto-decoders from single images. The mapping network consists of 3 convolutional layers with a kernel size of 5, whereas the mapping network is not required for datasets of video sequences or multi-view images such as UBCFashion, RenderPeople, and THUman2.0.

2.2 Experimental Results on DeepFashion

Qualitative and Quantitative Results. Existing auto-decoder-based methods generally require multi-view images or video sequences of objects to train an auto-decoder [14]. Instead, with the structured latent representation, we show that it is even possible to auto-decode 3D humans from single images on DeepFashion [11], as shown in Fig. 3, where our method generates realistic human images with reasonable geometry reconstructions (e.g., normal, depth). The quantitative results are shown in Tab. 1. Our method outperforms existing 3D GAN methods, including StyleSDF [15], EG3D [1], ENARF-GAN [13], and achieve comparable results as the publicly released EVA3D. However, the best performances of AG3D and EVA3D achieve lower FID.

Table 1: Quantitative results on DeepFashion. For reference, we report the quantitative results from the EVA3D (marked by '*') and the AG3D paper (marked by '∘').

FID ↓	DeepFashion
StyleSDF* [15]	92.40
EG3D* [1]	26.38
ENARF-GAN* [13]	77.03
EVA3D* [6]	15.91
EVA3D(Public)° [6]	20.45
AG3D° [2]	10.93
Ours	20.82

2.3 Editing in-the-wild images.

As shown in Fig. 4, to edit in-the-wild (cross-dataset) Internet images ①, an inversion step ② is performed, and images are edited via part-aware diffusion: ③ a new pose rendering, ④ new identity, ⑤ shoes, ⑥ T-shirts, ⑦ pants. Part-aware editing can also be applied for generated humans ⑧⑨, i.e., transferring the style of the Internet images (T-shirt ⑧), and editing the shoes ⑨.

2.4 Efficiency

Similar to most latent diffusion models *e.g.*, Stable Diffusion [16], ours is not trained end-to-end. It takes about 3.5 days to train an auto-decoder from about 80K images on UBCFashion and 3 days for latent diffusion on 4 NVIDIA V100 GPUs, more efficient than EVA3D [6] (5 days on 8 V100). In inference, our rendering network runs at 9.17 FPS to render 512^2 resolution images on a V100 GPU, 1.94× faster than AG3D. It takes about 127.55 s to sample 64 latents using DDIM [19] with 100 steps on one V100 GPU. Once trained, a new video sample can be added by inversion as shown in Fig. 5. Source video 1 and 2, with 110 frames each, take 1.75 min and 7.01 min (for complexity in hair) on 8 V100 to inverse respectively.

3 Limitation and Future Work

Limitation. **1)** We train models from scratch as in EVA3D/AG3D/ PrimDiff. The lack of a diverse in-the-wild human dataset with accurate registration is a common problem in this field. Due to the limited scale and dataset bias, diversity is not comparable to 2D diffusion models [16]. However, we outperform the baselines EVA3D and AG3D in diversity. **2)** Limited by the auto-decoder training, it is challenging to learn from single-view 2D image collections [14], *e.g.*, DeepFashion [11]. However, the structured latent representation makes it possible to auto-decode 3D humans from single images on DeepFashion, as shown

Fig. 4: Editing in-the-wild images. To edit in-the-wild images ①, we first apply inversion ②, and edit the images via part-aware diffusion.

in Fig. 3, where our method generates realistic human images with reasonable geometry reconstructions (*e.g.*, normal, depth).

Future Work. As discussed in Sec. 2.2, our framework is capable of auto-decoding 3D human latents from single images, and future work would be to improve the performances on DeepFashion.

References

1. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16123–16133 (2022)
2. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. ArXiv **abs/2305.02312** (2023), `https://api.semanticscholar.org/CorpusID:258461509`
3. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis (2020)
4. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. arXiv preprint arXiv:2002.10099 (2020)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015), `https://api.semanticscholar.org/CorpusID:206594692`

Fig. 5: Video inversion efficiency.

6. Hong, F., Chen, Z., Lan, Y., Pan, L., Liu, Z.: Eva3d: Compositional 3d human generation from 2d image collections. ArXiv **abs/2210.04888** (2022), https://api.semanticscholar.org/CorpusID:252780848
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR pp. 5967–5976 (2017)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. ArXiv **abs/1603.08155** (2016), https://api.semanticscholar.org/CorpusID:980236
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
10. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. CVPR pp. 6738–6746 (2017)
11. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR. pp. 1096–1104 (2016)
12. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
13. Noguchi, A., Sun, X., Lin, S., Harada, T.: Unsupervised learning of efficient geometry-aware neural articulated representations. arXiv preprint arXiv:2204.08839 (2022)
14. Ntavelis, E., Siarohin, A., Olszewski, K., Wang, C., Gool, L.V., Tulyakov, S.: Autodecoding latent 3d diffusion models (2023)
15. Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J.J., Kemelmacher-Shlizerman, I.: 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)
16. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2015)
18. Sitzmann, V., Martel, J.N.P., Bergman, A.W., Lindell, D.B., Wetzstein, G.: Implicit neural representations with periodic activation functions. ArXiv

Fig. 6: Qualitative results on RenderPeople. For our method, normal and depth maps at 128×64 resolution are rendered by our volumetric renderer.

abs/2006.09661 (2020), https://api.semanticscholar.org/CorpusID:219720931
19. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
20. https://renderpeople.com/3d-people/: Renderpeople (2018), https://renderpeople.com/3d-people/
21. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
22. Yu, T., Zheng, Z., Guo, K., Liu, P., Dai, Q., Liu, Y.: Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021) (June 2021)
23. Yue, Z., Loy, C.C.: Diffface: Blind face restoration with diffused error contraction. ArXiv **abs/2212.06512** (2022), https://api.semanticscholar.org/CorpusID:254591838
24. Zablotskaia, P., Siarohin, A., Zhao, B., Sigal, L.: Dwnet: Dense warp-based network for pose-guided human video generation. arXiv preprint arXiv:1910.09139 (2019)