3D-Aware Semantic-Guided Generative Model for Human Synthesis –Supplementary Material–

Jichao Zhang[1], Enver Sangineto[2], Hao Tang[3], Aliaksandr Siarohin[1,4], Zhun Zhong[1], Nicu Sebe[1], and Wei Wang[1]

[1]University of Trento [2]University of Modena and Reggio Emilia [3]ETH Zurich [4]Snap Research

Abstract. This document supplements our paper 3D-Aware GAN by providing more details of the implementation, training, the proposed metric as well as the additional experimental results on DeepFashion and another VITON dataset. Finally, we provide the detailed discussion about the training strategy and limitations of the model.

1 Implementation details

Training. Besides the training details reported in Sec. 5 of the paper, here we add that, in order to stabilize the training of G_{3D}, the loss weight λ_2 controlling the influence of ℓ_{ps} (Sec. 4.4 of the main paper) is set to 0 for the initial 100,000 training steps.

Dataset-specific parameters. The field of view is 10° for both datasets (DeepFashion and VITON). The camera elevation is 10° on both datasets, and the object rotation is 360° for DeepFashion and 72° for VITON. We estimated these parameters from the empirical distribution of each dataset. More details about the parameters and the network architecture can be found in our code. For the baselines, *i.e.,* GRAF [9], pi-GAN [2], GIRAFFE [6], ShadeGAN [7] and CIPS-3D [10], we used their publicly available code and we trained all the models using the architecture and the configuration corresponding to CelebA dataset [5]. The details of the CelebA configuration can be found in corresponding paper or in the public code of each baseline. However, for a fair comparison, we changed a few baseline parameters, such as the field of view, to be consistent with our model.

Average Matched Points (aMP). And we propose the average Matched Points (aMP) to evaluate the 3D-view consistency of the generated images based on local region matching. Specifically, we use Patch2Pix [11] to compute a point-wise matching between two generated images $(\tilde{I}_1, \tilde{I}_2)$ of the same person with different viewpoints, then we count the number of Matched Points $MP(\tilde{I}_1, \tilde{I}_2)$. MP is applied to image pairs with the same identity (texture) but different rotation angles. In more detail, for each method, we randomly generate 500 samples by varying the texture content while keeping fixed the other variation factors. Then, for each of these 500 samples, we change the camera viewpoint in order to get 3 different random rotations. We can now compute $MP(\tilde{I}_1, \tilde{I}_2)$ for all the 6 possible pairwise combinations of these 3 samples (note that

applying Patch2Pix to $(\tilde{I}_1, \tilde{I}_2)$ gets slightly different results from $(\tilde{I}_2, \tilde{I}_1)$). Finally, we average MP over the 500×6 pairs and we get a score which we call *average MP score* (aMP).

2 Additional results

The VITON dataset. In addition to the DeepFashion dataset [4], used in the main paper, we also compare our method with the baselines on the VITON dataset [3]. VITON is composed of 16,253 front-view woman and top-body clothes image pairs and it is widely used for virtual try-on tasks. We use the front-view woman images and we divide the dataset into 14,221 training images and 2,032 testing images. The original image resolution is 176×256, but we resize all the images to 256×256.

Unconditioned human image generation. Fig. 1 and Fig. 2 show a qualitative comparison between image samples generated by all the models using both the datasets. Our method generates more realistic human images than all the other baselines on both datasets. Note that our model achieves results comparable with HumanGAN [8] in human generation on the DeepFashion dataset. However, HumanGAN is not a 3D-Aware model, and it fails to control 3D factors. Note that HumanGAN [8] takes a source sample as input, from which most of the target appearance can be copied in the training stage. Conversely, our images are generated from noise, which is an harder task. Moreover, despite HumanGAN is trained on DeepFashion, there is no public code for training, thus for our comparison we had to use the sample images available in the official Web page (which are not generated unconditionally from noise, so probably the above comparison is a bit in favor of HumanGAN).

The quantitative evaluation provided in Tab. 1 Tab. 2 using the user studies and FID scores, shows that our approach significantly outperforms all the other methods also on the DeepFashion and VITON dataset. Fig. 8 and Fig. 9 show additional image generation results on both datasets, obtained using our method.

Table 1: User studies on the DeepFashion dataset.

Method	3D-SGAN	HumanGAN	GIRAFFE	CIPS-3D	pi-GAN	ShadeGAN	GRAF
User Study ↑	**55.2** %	41.2%	3.2%	0.2%	0.05%	0.0%	0.0%

Controllable human generation. Fig. 3 and Fig. 4 show a qualitative comparison using controllable human generation where we interpolate the "Rotation" parameter. Our method generates more realistic and more view-consistent results than the baselines. Moreover, Fig. 3 shows that GIRAFFE struggles to generate human images with different viewpoints. This is probably due to the fact that the degree for the field of view is small, and GIRAFFE may neglect the camera code and make the pose code learn most of the variations of the human body. Additional controllable human generation results are shown in Fig. 5 and Fig. 6.

Table 2: A quantitative comparison using FID scores on the VITON dataset.

Method	GRAF [9]	pi-GAN [2]	GIRAFFE [6]	ShadeGAN [7]	CIPS-3D [10]	3D-SGAN (Ours)
FID ↓	67.300	121.15	26.750	110.02	48.919	**14.060**

Real human image editing. Fig. 7 shows real human image reconstruction and editing results (see the main paper for the methodological details). Our model can reconstruct the real data with only minor differences with respect to the reference image (Fig. 7, Column 2). The other columns of Fig. 7 show the results of our method using latent code interpolations. Specifically, in top block of rows, we show semantics editing, in the second block, pose interpolation, and in last block, rotation editing. All the interpolations preserve the identity of the generated persons.

Visualization of the learning geometry.

Our 3D generator is similar to GIRAFFE, from which we adopted the low resolution feature rendering at 16×16, thus also our 3D predictions are not very informative. As shown in Fig. 10, we render the coarse normals from the density values.

3 Discussion

3.1 Separate training vs. joint training

As described in the main paper, G_{3D} and G_t are trained separately using the segmentation tensors as a bridge between the two generators. Note that a different solution, in which G_{3D} and G_t are jointly trained e.g., using adversarial learning, is possible. However, in our preliminary results, this solution led to significantly lower FID scores. Moreover, a solution in which the same generator is in charge of modeling both the 3D structure and the texture of the data, is conceptually similar to GIRAFFE, whose results are significantly inferior to our proposal in this full-body generation task. We presume the reason is that both DeepFashion and VITON are relatively small and GIRAFFE struggles to learn all the variation factors, including the human pose distribution, etc. (see Sec. 1 of the main paper). In contrast, 3D-SGAN splits the problem and the architectural design in two stages. The advantages are that, this way, we can: (1) simplify the learning problem, (2) use ground truth segmentation masks (automatically obtained using [1]) as an additional supervision (e.g., used in ℓ_r). On the other hand, a potential disadvantage in separate training is a possible domain gap between training and inference, since G_t, at inference time, is fed with segmentation tensors generated by G_{3D} which have not been observed at training time. Despite that, our empirical results show that G_t is robust enough to this domain shift.

3.2 View-inconsistency

The largest limitation of our method is the lack of a full 3D consistency of the generated textures with respect to multiple views of the same person. However, this problem is shared by most 3D-aware GANs, because training does not include multiple

Fig. 1: Unconditioned human image generation. A comparison between our 3D-SGAN with all the baselines using the VITON dataset.

paired views of the same person or other similar supervision. Moreover, we inherit from GIRAFFE (whose structure is adopted in our G_{3D}) the "mirror symmetry" problem [10], which depends on the way point coordinates are represented ($\gamma(\boldsymbol{x})$). Nevertheless, 3D-SGAN can alleviate the consistency issues to a large extent. Specifically, when we generate multiple views of the same person, we change the camera pose $\boldsymbol{z}_c$ *keeping fixed* all the other latent codes. In particular, $\boldsymbol{z}_t$ is *fixed* and it is sampled *only once* when, e.g., we produce the "Rotation" results. This way we can generate multiview images of the same person with *the same* overall texture (e.g., for the clothes). Moreover, 3D consistency is further encouraged by the proposed consistency losses. Although we do not fully solve the problem (e.g., we cannot control the face details), nevertheless we can alleviate it, as empirically shown by the comparison with respect to the baselines.

Fig. 2: Unconditioned human image generation. A comparison between our 3D-SGAN with all the baselines using the DeepFashion dataset.

Fig. 3: VITON dataset: controllable image synthesis by "Rotation" interpolation.

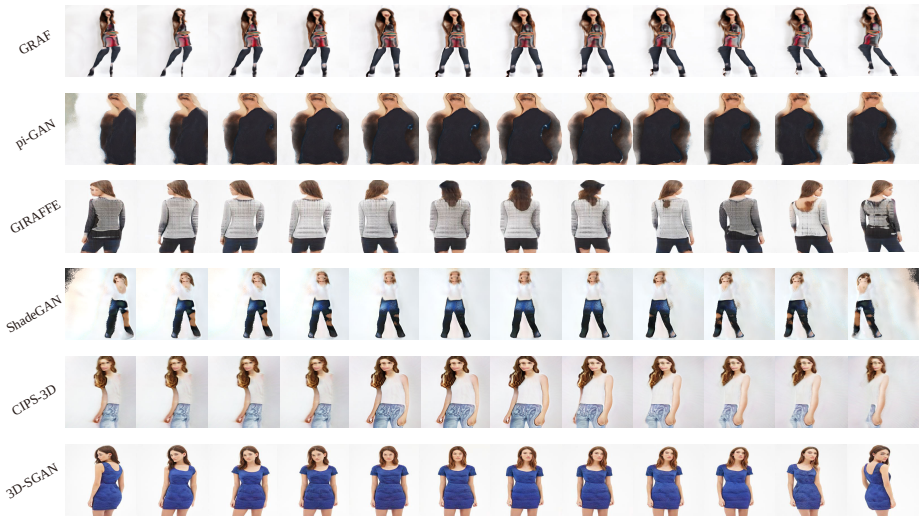

Fig. 4: DeepFashion dataset: controllable image synthesis by "Rotation" interpolation.

Fig. 5: VITON dataset: controllable person generation by interpolating different latent codes: 'Object Pose', 'Semantics', 'Texture', 'Translation'. For 'Translation', we show generation results for 'Horizontal Translation', 'Vertical Translation', and 'Depth Translation', respectively.

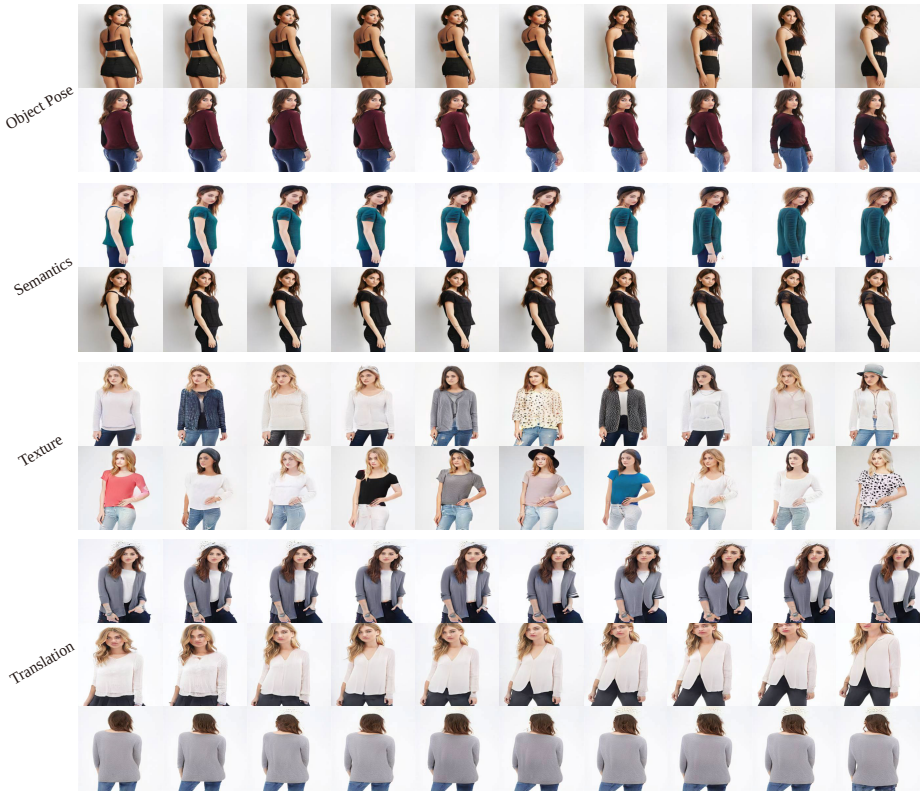

Fig. 6: DeepFashion dataset: controllable person generation by interpolating different latent codes: 'Object Pose', 'Semantics', 'Texture', 'Translation'. For 'Translation', we show generation results for 'Horizontal Translation', 'Vertical Translation', and 'Depth Translation', respectively.

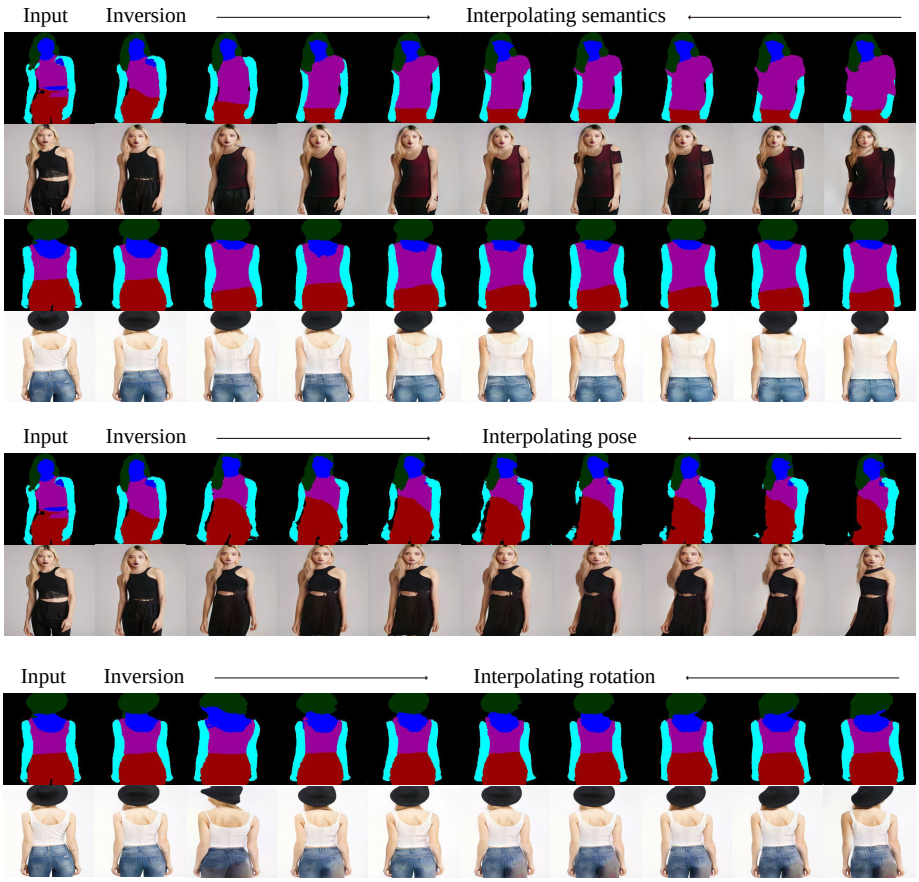

Fig. 7: DeepFashion testing dataset: real data reconstruction and editing.

Fig. 8: VITON dataset: additional randomly generated images using our method.

Fig. 9: Deepfashion dataset: additional randomly generated images using our method.

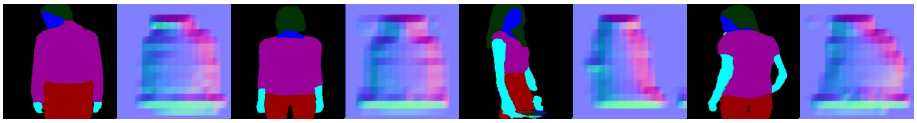

Fig. 10: The visualization of 3D features using the normals.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39** (2017) 3
2. Chan, E., Monteiro, M., Kellnhofer, P., Wu, J., Wetzstein, G.: pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In: CVPR (2021) 1, 3
3. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: CVPR (2018) 2
4. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: CVPR (2016) 2
5. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. 2015 IEEE International Conference on Computer Vision (ICCV) pp. 3730–3738 (2015) 1
6. Niemeyer, M., Geiger, A.: GIRAFFE: Representing scenes as compositional generative neural feature fields. In: CVPR (2021) 1, 3
7. Pan, X., Xu, X., Loy, C.C., Theobalt, C., Dai, B.: A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. NeurIPS (2021) 1, 3
8. Sarkar, K., Liu, L., Golyanik, V., Theobalt, C.: Humangan: A generative model of humans images (2021) 2
9. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. In: NeurIPS (2020) 1, 3
10. Zhou, P., Xie, L., Ni, B., Tian, Q.: CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis (2021) 1, 3, 4
11. Zhou, Q., Sattler, T., Leal-Taixe, L.: Patch2pix: Epipolar-guided pixel-level correspondences. In: CVPR (2021) 1