

3DToonify: Creating Your High-Fidelity 3D Stylized Avatar Easily from 2D Portrait Images

(Supplementary Material)

1. Video demonstration of 3DToonify

Given a portrait video captured by a monocular camera, our goal is to reconstruct and stylize the underlying 3D structures following a single target exemplar to generate high-fidelity 3D stylized avatars. Previous 2D portrait stylization methods [2, 5] can not maintain 3D view-consistency under extreme viewpoints, and existing 3D avatar stylization methods [3, 6] also fail to achieve high-fidelity 360° stylization with user-specific styles, as demonstrated in part I of the supplementary video (file ‘5342_video.mp4’). To this end, we propose to utilize the spatial neural representation (SNR), implicit functions with spatial-shared attributes, to capture information in the 3D space, and introduce a progressive training scheme for learning stylized and disentangled structures (i.e., deformed geometry and stylized texture). We provide visualization of this scheme in part II. In part III, we show multiple stylized portraits rendered from arbitrary novel viewpoints with 3D consistency. With the implicit representation built by SNR, we can easily infer stylized results in interpolated novel views via volume rendering. Furthermore, our method enables explicit 3D assets export, thus allowing for real-time rendering and arbitrary-view stylization with flexible user control. In part IV, we compare multi-view rendering results in stage II and stage III, where the latter one shows more stable results with 3D consistency, demonstrating the effectiveness of the proposed decomposed texture field in learning spatial-shared information in the entire 3D space.

2. Architecture of MVS-guided prior learning

In the proposed photorealistic prior learning stage, except for the commonly-used radiance color constraints from 2D image observations, we also extract depth maps from the captured photos leveraging multi-view stereo (MVS) methods and use them as extra geometric guidance by adding the depth loss and the surface loss. The overview flowchart of this part is shown in Figure 1.

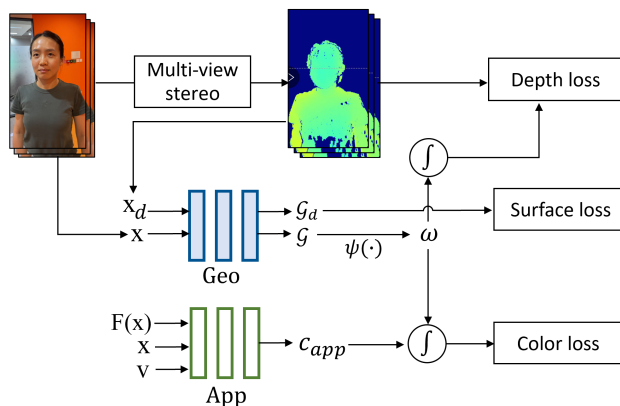


Figure 1. The architecture of MVS guided prior learning.

3. Impact of the number of stylized frames

The stage II and III of our method aim to adapt the underlying structures (e.g., geometry and texture) of the stylized avatar with the supervision of a set of few-shot 2D stylized portraits. In order to avoid the style inconsistency brought by existing 2D portrait stylization methods on side-face style transfer, we only use forward/backward stylizations within a small yaw angle (0.2 radian) as supervision for the style adaption process, where the yaw angle can be automatically estimated from facial landmarks. Hence, the number of stylized training frames may not be a fixed value for each scene and depend on the duration of the moving camera from different viewpoints (commonly around 50-100 in a dense sampling). To validate the impact brought by the number of stylized frames used for supervision, we train our model on video ‘‘Woman1’’ with 80, 40, 8, 4, 1 frames respectively (from automatically computed frames to uniformly-sampled ones). As shown in Figure 2, the geometry adaption is robust to the number of few-shot stylizations. Our model trained on 1 frame still enables a deformed surface from the original prior. Table 1 shows the quantitative results, which indicate that training on sufficient frames improves the performance of stylized avatar

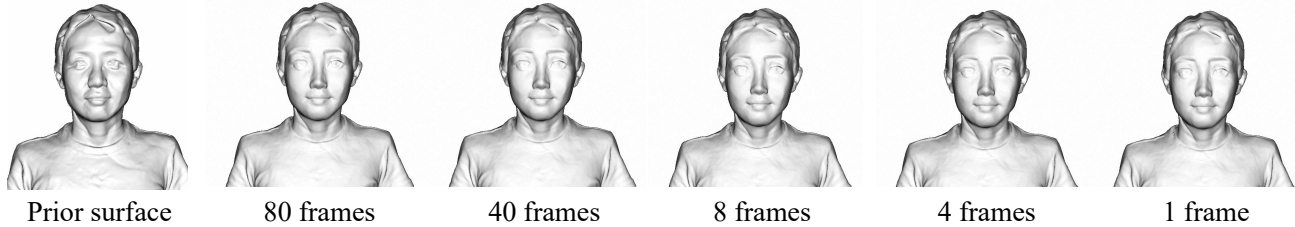


Figure 2. Geometry results of model trained with different number of stylize training frames.

Table 1. Quantitative results of model trained with different number of stylize training frames.

Frames	1	4	8	40	80
FID ↓	98.3	96.2	95.8	96.2	96.7

synthesis. However, training on too many frames may decrease the performance, as the network has difficulty fitting all the information.



Figure 3. Geometry comparison before and after stylization.

4. Geometry comparisons before and after stylization

To demonstrate the effectiveness of geometry adaption, we provide geometry comparison results before and after stylization for different cases shown in the main paper. As illustrated in Figure 3, the geometry adaption stage enables disentangled learning of geometry and appearance, thus allowing the underlying geometry stay faithful to the external styles.

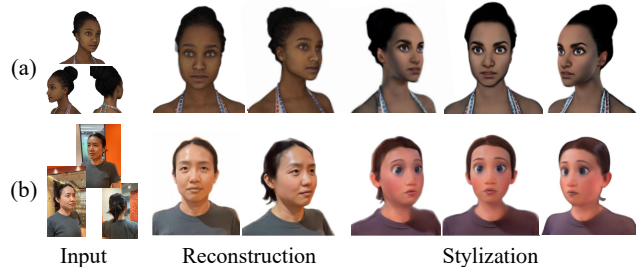


Figure 4. More diverse results.

5. Diverse stylized results in race and lighting

We provide additional visual results for different races and extreme lighting conditions. For race, we show an example of a dark-skinned woman in Figure 4(a), which demonstrates that our method can handle various skin tones. For lighting, we show results on a captured video under complicated lighting conditions (e.g., obvious reflections on forehead and hair, shadows on clothing) in Figure 4(b). Reasonable results are stably produced with the help of the subtle progressive training scheme. The shown results demonstrate the effectiveness of our method when handling diverse and complicated inputs.

6. Results of explicit stylized avatar

After training, explicit 3D models can be easily extracted from the spatial-shared representation learned in SNR. Geometric surface is computed as the zero-set of SDF and the Marching Cubes algorithm [1] is used to extract connective faces. Vertex colors can be inferred by accumulating colors from the texture field with the closest preset view, which are then converted to atlas textures by the cylinder UV mapping. The extracted explicit 3D model enables flexible user control and editing (e.g., you can feed the texture map into 2D style translator to further enhance the stylization effects.), which are compatible in existing modern graphic workflow. Demonstration of the explicit stylized avatar is shown in Figure 5.






Exemplar	Corresponding text description
	<i>Pixar style, a young girl with long, curly black hair and brown skin, high-quality, cartoon characters, realistic animation.</i>
	<i>Comic book style, comic style, gray skin, dark-skinned, black hair, blue background, serious expression on his face.</i>
	<i>Disney style, Rapunzel, Disney princess in Tangled, cartoonish and whimsical, bright, vibrant colors.</i>
	<i>Disney style, princess Fiona in Shrek, green skin and red hair, cartoonish and whimsical.</i>
	<i>Disney style, Hiccup Horrendous Haddock III in How To Train Your Dragon, cartoonish and whimsical, bright, vibrant colors.</i>

Table 2. Text descriptions of corresponding target style exemplars used as input for NeRF-Art [3].

7. Text descriptions of target style exemplars

Here we list the text prompts used to describe the target style exemplars in Table 2, which serve as the input for NeRF-Art [3] in qualitative comparisons conducted in Section 4.1. The exemplars are selected from the training dataset of DualStyleGAN [4], and the corresponding text prompts are generated by Mini-GPT4 [7].

8. Limitations

Stylized texture recovery. Our method can disentangle the textural colors with spatial-shared attributes, but absolute albedo extraction is still limited to the situation of uniform illumination. Due to the difficulty of the forward renderer to simulate realistic real-world scenarios with complicated light condition, it is more challenging and ambiguous for inverse rendering to recover the original 3D materials. Therefore, the final stylized albedo texture obtained by our method is only a reasonable 3D-consistent stylization on the original reconstructed contents, which means that even though the specular reflection varying with moving viewpoints can be successfully removed by the spatial-shared texture field, the shadow caused by uneven illumination distribution will still remain. However, it is possible to further refine the texture directly on the exported 3D assets, since explicit models can be easily extracted from the spatial-shared representation, which also allows for more flexible user control.

Static human scene. Our method is proposed for static human scene and the performer should stay still during the video capture. Dynamic portraits may bring confused surface and blurred texture, due to the misalignment between multi-view observations on pixel-level.

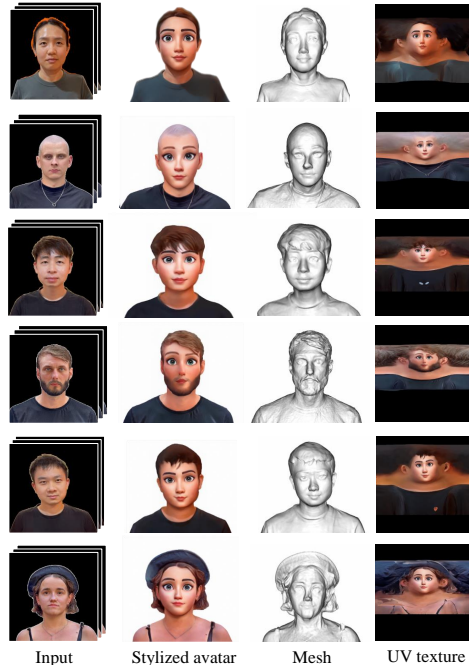


Figure 5. Results of explicit stylized avatar. Given the capture of 2D portrait images, we can further obtain explicit stylized avatars with geometric mesh and UV texture.

References

- [1] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 2
- [2] Yifang Men, Yuan Yao, Miaomiao Cui, Zhouhui Lian, and Xuansong Xie. Dct-net: domain-calibrated translation for portrait stylization. *ACM Transactions on Graphics (TOG)*, 41(4):1–9, 2022. 1
- [3] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 3
- [4] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7693–7702, 2022. 3
- [5] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Vtoonify: Controllable high-resolution portrait video style transfer. *arXiv preprint arXiv:2209.11224*, 2022. 1
- [6] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9144–9154, 2023. 1
- [7] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3