Image-Guided Geometric Stylization of 3D Meshes

Supplementary Material

A. Implementation Details

To train LoRA weight [3] on Stable Diffusion XL [10] using DreamBooth [12], we use a total of 4-12 images for each style reference and set the LoRA rank to 16. Following common practice, we adopt the prompt "A photo of TOK sculpture" and omit the class-specific prior preservation loss proposed in the original DreamBooth formulation. The learning rate is set to 10^{-4}, and LoRA weight is trained for a total of 800 iterations. In the case of the approximated VAE [6] encoder, we first render $N = 500$ images of the source mesh from random viewpoints and obtain corresponding latent via SDXL VAE encoder. We then fit the matrix using these 500 image–latent pairs.

For mesh deformation, the SDS loss [11] is computed using the text prompt "A TOK style sculpture". We use $N_1 = 1800$ and $N_2 = 3600$ iterations for optimization. Rasterization is performed with the differentiable nvdiffrast rasterizer [7]. The camera FOV and distance are set to $30°$ and 5.0, respectively. The viewpoint is randomly sampled by choosing the elevation uniformly from $[10°, 30°]$ and the azimuth from $[0°, 360°)$. Batch size is set to 4, *i.e* We render mesh from four different viewpoint at each iteration. We optionally leverage the proposed symmetry loss, when the symmetry is detected via Eq. (8) of the main paper and the source mesh's vertices are arranged accordingly.

B. Details of User Study

We quantitatively compare the proposed method with baselines through a perceptual user study. A total of 32 participants were gathered and each participant was requested to rank the output of a total of six different methods [1, 4, 5, 8, 9] for 8 samples based on three criteria:

- **1) Geometric Alignment**: Rank the 3D model based on how accurately it aligns with the reference image in terms of geometric properties (pose, silhouette, and structural shapes). (A higher rank indicates closer geometric correspondence to the reference image.)
- **2) Content Preservation**: Rank the 3D model based on how well it maintains characteristics of the original 3D model, even after incorporating the reference image's style. (A higher rank means the 3D model remains more faithful to the original model while integrating the new style.)
- **3) Aesthetic Style Consistency**: Rank the 3D model based on how well it captures the overall artistic style or visual impression of the reference image. (A higher rank means the model conveys a style that feels more consistent with the reference.)

For each question, we randomly shuffle the order of the outputs from six different methods to ensure a fair comparison.

C. Additional Results

We visualize the additional qualitative comparison with the baselines [1, 4, 8, 9] in Fig. 11. As shown, our method effectively deforms the source mesh to incorporate the geometric style of the style reference, while baseline methods tend to generate artifacts or only texture-like small modifications. We show the additional qualitative results in Fig. 12, highlighting the superior performance of the proposed method.

In addition, we report the FID score [2] between rendered meshes and style references for additional quantitative evaluation. Each deformed mesh is rendered from 16 different viewpoints, and the corresponding style images employed during LoRA [3] training are used as the reference set. A total of 12 meshes are used for evaluation. As reported in Table 1, our method achieves the best score, demonstrating superior perceptual alignment with the target style compared to baselines.

Method	FID (↓)
Paparazzi [8]	419.89
Neural 3D Mesh Renderer [4]	417.08
MeshUp [5]	396.23
Text2Mesh [9]	409.17
TextDeformer [1]	400.58
Ours	**376.57**

Table 1. Quantitative comparison measured by FID [2].

References

[1] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *SIGGRAPH*, 2023. 1, 2

[2] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 2017. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 1

[4] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 1, 2

[5] Hyunwoo Kim, Itai Lang, Noam Aigerman, Thibault Groueix, Vladimir G Kim, and Rana Hanocka. Meshup: Multi-target mesh deformation via blended score distillation. In *3DV*, 2025. 1, 2

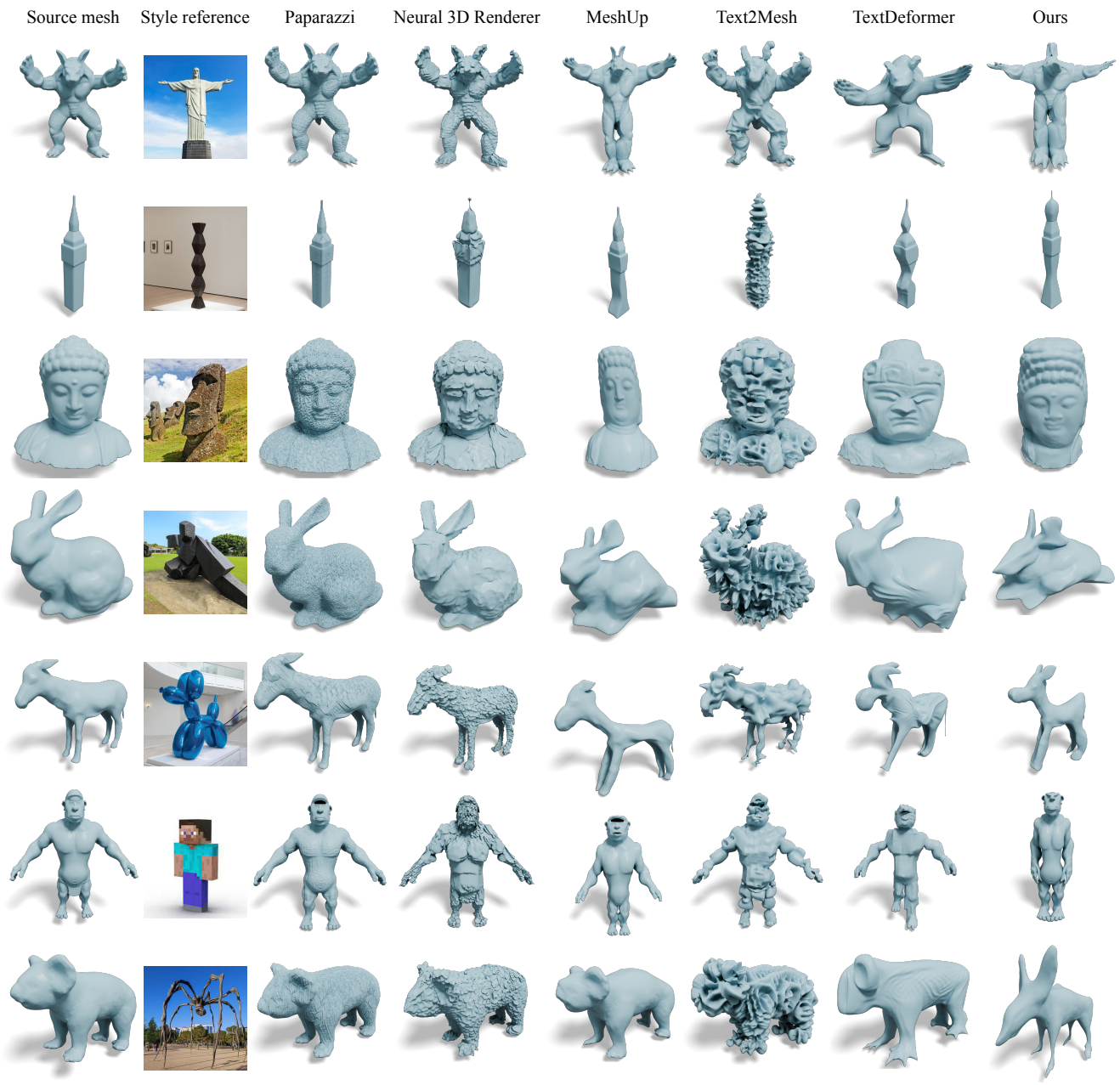

Figure 11. Additional qualitative comparison of our method with the baselines [1, 4, 5, 8, 9] on diverse geometric stylization scenarios.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2013. 1

[7] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM TOG*, 2020. 1

[8] Hsueh-Ti Derek Liu, Michael Tao, and Alec Jacobson. Paparazzi: surface editing by way of multi-view image processing. *ACM TOG*, 2018. 1, 2

[9] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *CVPR*, 2022. 1, 2

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023. 1

[11] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 1

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch,

Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

Figure 12. Additional qualitative results of the proposed method.