Compress3D: a Compressed Latent Space for 3D Generation from a Single Image (Supplementary Material)

Bowen Zhang[1*], Tianyu Yang[2†], Yu Li[2], Lei Zhang[2], and Xi Zhao[1†]

[1] Xi'an Jiaotong University
[2] International Digital Economy Academy (IDEA)

1 Network Details

In this section, we present network details of the diffusion prior model and triplane diffusion model.

Diffusion Prior Model The network architecture of the diffusion prior model is a MLP with skip connections between layers at different depths as shown in Fig. 1. The diffusion backbone consists of multiple ResBlocks. In each block, image embedding is injected into the block through concatenation, and the timestep embedding is injected through addition.

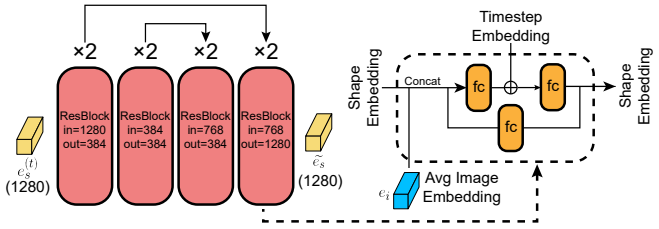

Fig. 1: Diffusion Prior Network.

Triplane Diffusion Model The diffusion backbone is a UNet, which contains multiple ResBlocks and down/up sample layers as shown in Fig. 2. The input and output of each ResBlock are triplanes, we use TriConv (3D-aware convolution [3]) in each ResBlock, and the timestep embedding is injected through addition. Shape embedding e_s and image embedding e_p are injected through cross attention. Specifically, shape embedding e_s and image embedding e_p are concatenated, which are subsequently used for predicting $k \in \mathbb{R}^{l \times c}$ and $v \in \mathbb{R}^{l \times c}$, where l is the number of tokens, c is the channels number of each token. The $q \in \mathbb{R}^{r^2 \times c}$ is a reshaped triplane, r is the resolution of triplane. The cross-attention feature is obtained by $\texttt{sotfmax}(\frac{qk^T}{\sqrt{c}})v$, and finally we reshape the cross-attention feature back to a triplane.

* Work done during the internship at IDEA.
† Corresponding authors.

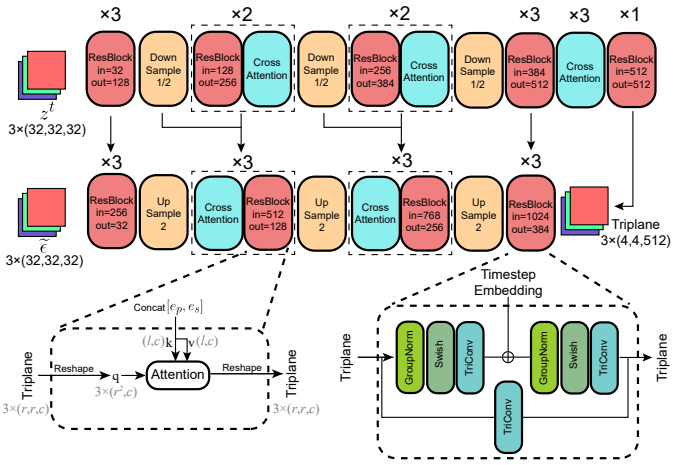

Fig. 2: Triplane Diffusion UNet.

2 Images In the Wild

In addition to testing our method on the test dataset, we also test our method on some images outside the dataset. Specifically, we collect some images on the internet as input. The generation results are shown in Fig. 3, demonstrating that our method generalizes well to images in the wild.

3 Limitations and Future Work

While Compress3D has demonstrated success in generating high-quality 3D models, there are certain limitations inherent in our approach. As illustrated in Fig.4, the resolution of 3D scene representation, as discussed in [2], imposes constraints on the level of detail achievable for thin structures in our generated results, which may not match those of manually modeled 3D models. Additionally, since Compress3D is trained on the Objaverse dataset [1], which includes samples with illumination information, the textures generated by our method also incorporate embedded illumination details. As depicted in the second example of Fig.4, the front of the stele appears lighter, while the back appears darker, with shadows cast on the ground. Furthermore, owing to the nature of generative models, our method may produce some 3D models that visually appear correct but deviate from reality. For instance, when provided with a photo of the Earth, our method may generate a planet resembling the Earth, yet it may not precisely replicate the real Earth. Looking ahead, our future efforts will be directed towards enhancing the texture quality of 3D models and providing greater control over the generation process.

4 More Results

We show more results of our method, in Fig. 5, Fig. 6, Fig. 7 and Fig. 8.

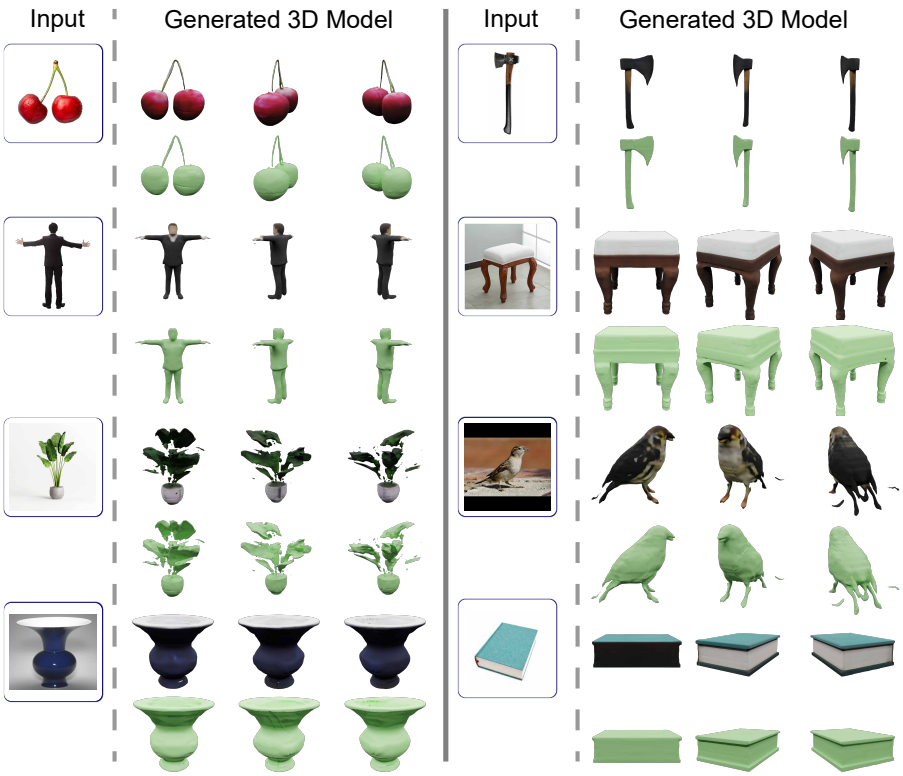

Fig. 3: Generation results for images in the wild.

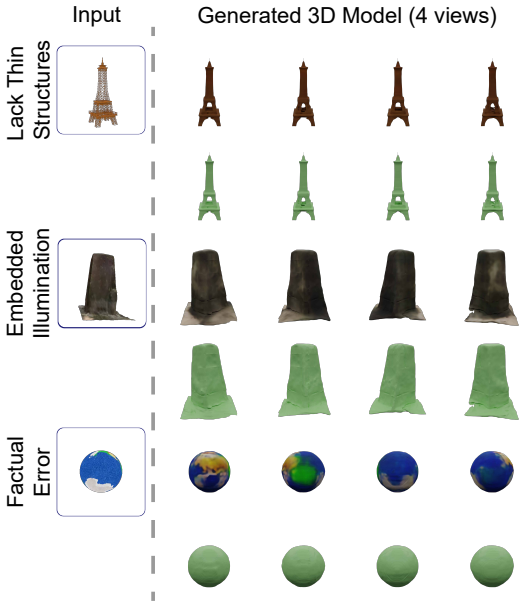

Fig. 4: Limitations of our method.

Fig. 5: The generation diversity of our method.

References

1. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
2. Shen, T., Munkberg, J., Hasselgren, J., Yin, K., Wang, Z., Chen, W., Gojcic, Z., Fidler, S., Sharp, N., Gao, J.: Flexible isosurface extraction for gradient-based mesh optimization. ACM Transactions on Graphics (TOG) **42**(4), 1–16 (2023)
3. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023)

Fig. 6: More results of our method.

Fig. 7: More results of our method.

Fig. 8: More results of our method.