Supplementary Material for: SF3D: Stable Fast 3D Mesh Reconstruction with UV-unwrapping and Illumination Disentanglement

Mark Boss[1] Zixuan Huang[1,2†] Aaryaman Vasishta[1] Varun Jampani[1]

[1]Stability AI [2]UIUC

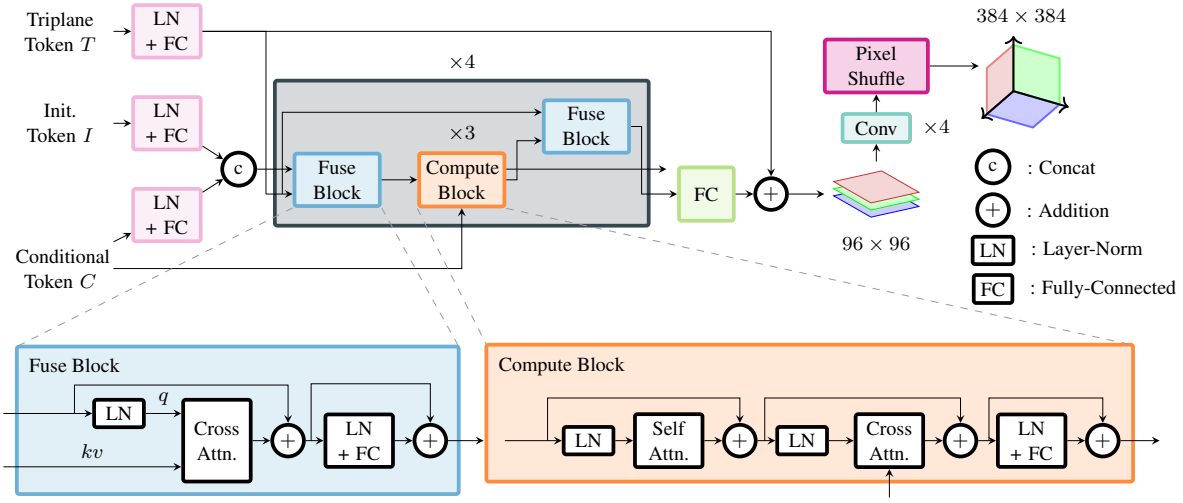

Figure 1. **Enhanced Transformer Architecture.** Our new backbone produces higher resolution output triplanes. We further upscale them using a pixel shuffling [2]. This helps capture high-frequency, high-contrast textures with reduced aliasing as in Fig. 4 of the main paper.

1. Enhanced Transformer

To reduce the aliasing artifact, we upgrade the transformer backbone to produce triplanes at a resolution of 384×384. However, naively increasing triplane tokens in TripoSR [3] is computationally prohibitive due to the quadratic complexity of self-attention. Inspired by PointInfinity [1], we leverage a two-stream transformer, which has linear complexity w.r.t. the number of tokens. As illustrated in Fig. 1, our architecture consists of two processing streams, the triplane stream and the latent stream. The triplane stream consists of the raw triplane tokens to be processed. In each two-stream unit (gray box in Fig. 1), the latent stream fetches information from the triplane stream using cross attention and performs the main computation on a set of constant-sized latent tokens. The latent stream then updates the triplane stream with the processed latent tokens. Our full architecture consists of four such two-stream units. With this computationally detached design, our transformer can produce triplanes at a resolution of 96×96 with 1024 channels. To further increase the resolution and reduce aliasing, we integrated a pixel shuffling operation [2], enhancing the triplane resolution to 384×384 with a feature dimension of 40.

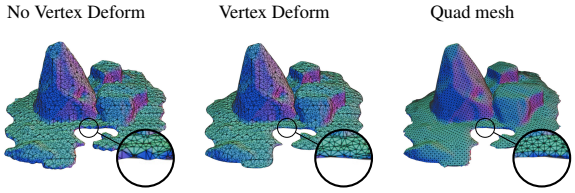

Figure 2. Visualization of topology and normal smoothness.

2. Influence on topology

In Fig. 2 we show the influence of our vertex deformation and additional influences of postprocess retopology. Notice the non-smooth highlighted sections and how the vertex deformation produces a smoother surface. The topology is still not ideal for downstream applications, but a simple quad remeshing can achieve good output topology which benefits from our smooth surfaces.

3. Additional Results

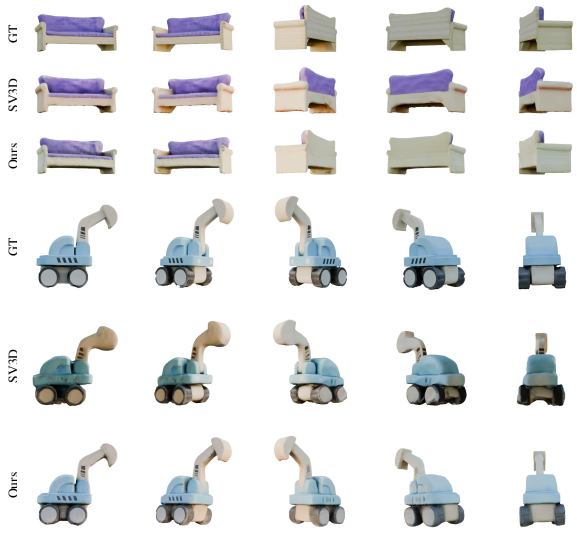

Figure 3. **Novel View Synthesis.** Even compared to SV3D, a SOTA diffusion-based method that takes 20 minutes per asset, our method produces more or similarly detailed assets that follow the geometry well.

We selected some of the publicly released meshes from SV3D [4] and compared them with our reconstructions in Fig. 3. Note that the SV3D conditioning images are not available directly. Hence, this only acts as a weak comparison. Not that our method is capable of producing a similar level of quality with a drastically lower inference budget.

In Fig. 4, we present further results on our decomposition task. Notice the plausible diffuse colors with severely reduce illumination bake-in and the overall metallic and roughness decomposition. When we perform relighting, the assets also create realistic light interactions and fit well in these novel illuminations.

In Fig. 5, we present further comparisons with all methods on GSO and OmniObject3D objects. Our method produces consistent and plausible reconstructions from these test datasets.

References

[1] Zixuan Huang, Justin Johnson, Shoubhik Debnath, James M Rehg, and Chao-Yuan Wu. Pointinfinity: Resolution-invariant point diffusion models. In *CVPR*, 2024. 1

[2] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *CVPR*, 2016. 1

[3] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3D object reconstruction from a single image. *arXiv*, 2024. 1

[4] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv*, 2024. 2

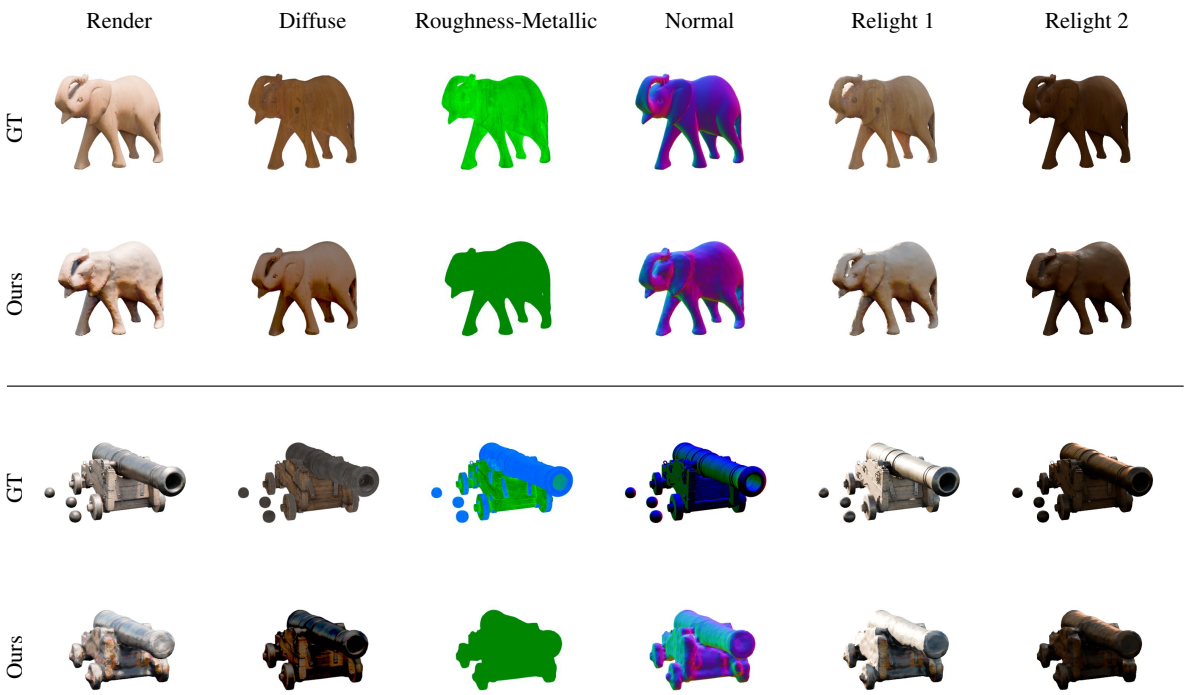

Figure 4. **Decomposition Results.** Additional results from high-quality objects. These are rendered under under natural illumination. These illuminations are highly challenging for material estimation. Still our model estimates sensible material properties, which allow for a convincing relighting.

Figure 5. **Comparison on GSO and OmniObject3D.** Notice how our reconstructions produce consistent results with detailed textures and smooth shading.