Supplementary Material

1. Qualitative Comparison on Generated Images and Real-World Images

We also provide a qualitative evaluation of novel view synthesis on real-world images, comparing our method with One-2-3-45 [2] and Zero-1-2-3 [3]. The real-world images used in our evaluation are sourced from One-2-3-45[1], captured in real-world or generated from 2D image generative model (e.g. DALL-E [4]). We leverage the elevation estimation module from One-2-3-45 to estimate the elevation of the input and assume the azimuth of the input is 0. In Figure 2, a qualitative comparison between ours and baselines is presented. Similar to the observed experimental phenomena on the GSO [1], Zero-1-2-3 faces challenges in maintaining consistency across different novel views. While One-2-3-45 demonstrates proficiency in geometry recovery, it exhibits limitations in the quality of rendering images. In contrast, our method not only achieves high-quality novel view synthesis but also maintains consistency across views, closely adhering to the input image. Despite being trained on a synthetic dataset, our method generalizes well to real-world images. Moreover, it exhibits robustness to estimated camera poses.

Figure 3 shows more results of our method on images generated by Midjourney and real-world images captured by the phone. It also demonstrates that our method has a good ability to generalization and maintain high-quality reconstruction for various objects.

2. More Qualitative results

We present additional visual comparisons between the two baselines and ours in Figure 4. Consistent with previous findings, Zero-1-2-3 struggles to maintain consistency while One-2-3-4-5 has relatively poor performance on novel view synthesis.

Furthermore, we have included an offline web demo in the supplementary materials to showcase the 360-degree visualizations of both the other baselines and our method. Please refer to it for more details.

3. Ablation on the influence of Mask, Distance Transform, and Camera Modulation to Geometry Reconstruction

The local feature is employed to enhance the Gaussian features and guide point cloud up-sampling. We only include mask and DT in the former, while excluding them in the latter. The shape is represented by point clouds, and triplanes further predict offsets for better 3D Gaussian rendering. We think the mask and DT do not provide more local shape information for point up-sampling, while they would help the triplane to refine some generated 'outside' points by predicted offset. Our primary focus for upsampling is to replace the global feature of coarse geometry (in the original SPD module) with more informative local image features (e.g., DINOv2). Table 1 shows the mask and DT do not enhance the shape output well. Moreover, camera modulation plays a pivotal role in results by providing viewpoint information. During testing, we obtain a mask using the 'rembg' tool and the SAM Model for background removal. This mask is applied to eliminate images with a clean background (e.g., white) and also provides extra local features for Gaussian feature enhancement.

Table 1. Ablation evaluation about the mask, distance transform (DT), and camera modulation.

	CD ↓	IoU ↑
Base	21.04	0.401
+ DT	22.10	0.391
+ Mask	21.71	0.393
+ Mask & DT	21.95	0.390
w/o Cam	37.43	0.362

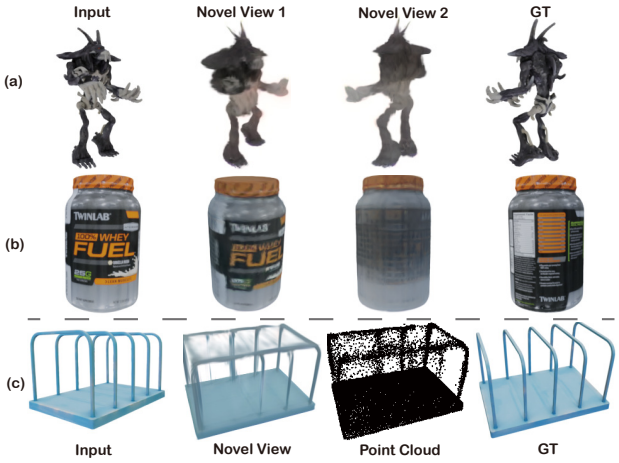

Figure 1. Failure cases.

4. Failure Cases and Limitations

While our method has demonstrated effectiveness, there are still some limitations, as illustrated in Figure 1. As dis-

[1] https://github.com/One-2-3-45/One-2-3-45/tree/master/demo/demo_examples

cussed in the paper, our regression method often struggles to "imagine" the backside (except it sometimes can guess the backside through shape symmetric prior as Figure **??** (c) shown), resulting in blurry texture (see Figure 1 (b)). While our method can roughly reconstruct the geometry, it encounters challenges in accurately reconstructing complex action figures, as depicted in Figure 1 (a). Furthermore, Figure 1 (c) illustrates that inaccurate point cloud estimation can adversely affect the accuracy of our 3D Gaussians. To improve our method, the potential solutions could include: (1) designing a mechanism to facilitate feature interactions between the point cloud decoder and the triplane decoder and (2) exploring a diffusion model based on 3D Gaussian to achieve improved texture results, especially on the opposite side.

References

[1] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas Barlow McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, pages 2553–2560, 2022. 1

[2] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 1

[3] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. 2023. 1

[4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 1

Figure 2. Qualitative comparison with Zero-1-2-3 and One-2-3-45 on real-world images.

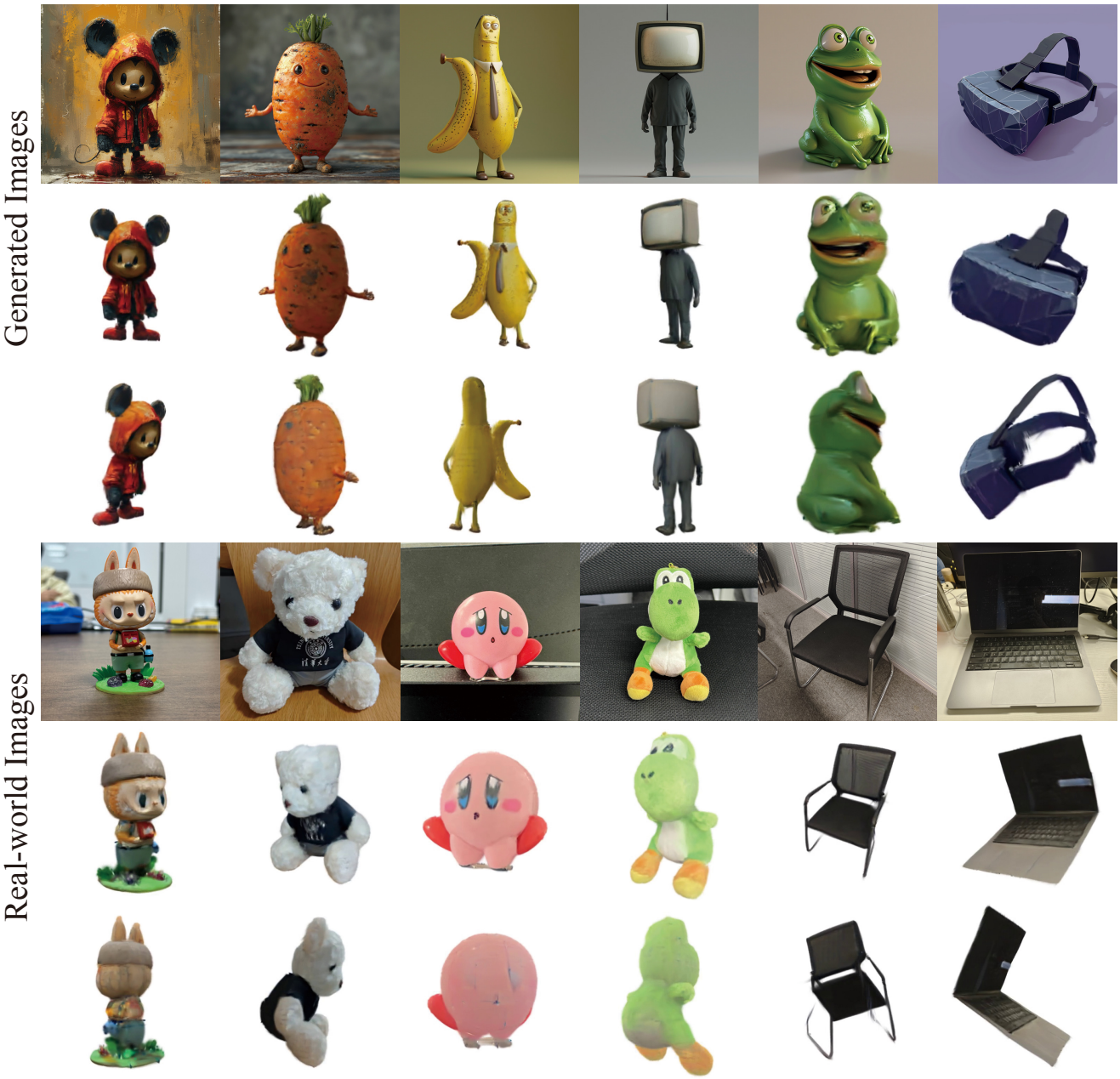

Figure 3. More qualitative results of our method on generated images (top) and real-world captured images (bottom).

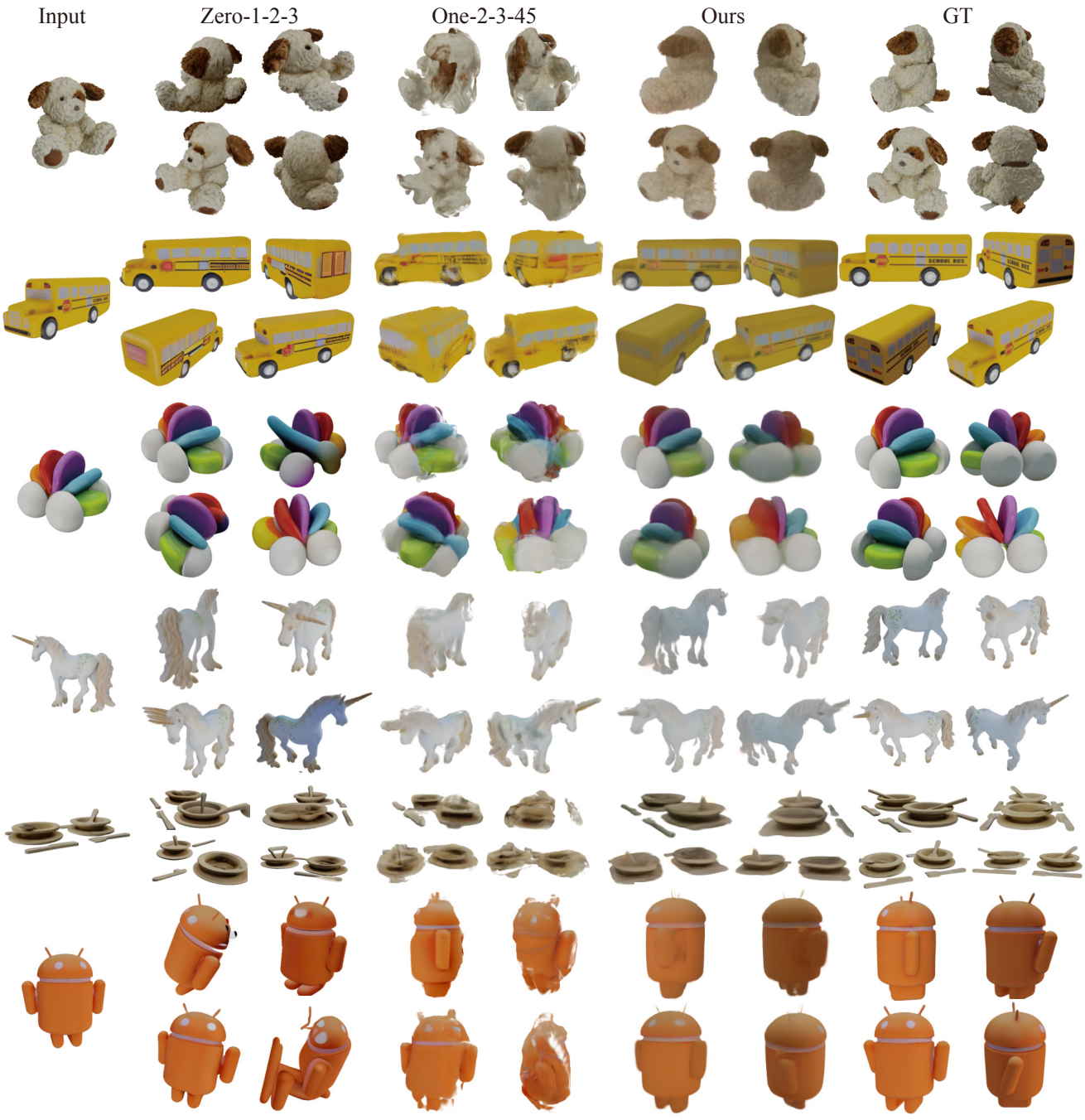

Figure 4. More qualitative comparison with Zero-1-2-3 and One-2-3-45 on GSO dataset.