High-fidelity 3D Object Generation from Single Image with RGBN-Volume Gaussian Reconstruction Model

Supplementary Material

1. Method Details

1.1. Regularization Loss Function

The regularization loss function [2] contains self-supervised depth distortion loss $\mathcal{L}_{\mathrm{d}}$ and normal consistency loss $\mathcal{L}_{\mathrm{n}}$. Drawing inspiration from Mip-NeRF [1], the depth distortion loss function focuses the weight distribution along the rays by minimizing the distance between ray-primitive intersections, which can be formulated as

$$\mathcal{L}_{\mathrm{d}} = \sum_{i,j} \omega_i \omega_j \left| z_i - z_j \right| \tag{1}$$

where $\omega_i = \alpha_i \mathcal{G}_i(\mathbf{u}(\mathbf{x})) \prod_{j=1}^{i-1} (1 - \alpha_j \mathcal{G}_j(\mathbf{u}(\mathbf{x})))$ denotes the blending weight of i-th intersection and z_i denotes the depth of the intersection point. 2D Gaussians explicitly model the primitive normals, enabling us to align their normals n_i with the normals N derived from depth maps through the normal consistency loss function, which can be formulated as

$$\mathcal{L}_{\mathrm{n}} = \sum_{i} \omega_i \left(1 - \mathbf{n}_i^{\top} \mathbf{N}\right) \tag{2}$$

1.2. Mesh Extraction

Following 2D Gaussian Splatting [2], We generate depth maps of the training views by extracting the depth values from the splats projected onto the pixels and employ truncated signed distance fusion (TSDF) with Open3D to integrate these reconstructed depth maps, leading to high-quality mesh extractions from generated 2D Gaussians. To ensure a fair comparison, we also employ the widely adopted mesh extraction technique proposed by LGM [6], which also yields equally satisfactory textured meshes.

2. Additional Qualitative Results

We present additional qualitative results of novel view synthesis and single view reconstruction, as depicted in Fig. 1, 2, and 3. Our method demonstrates the generation of high-quality 3D objects with consistent semantic and geometric details, surpassing the performance of baseline approaches. Furthermore, we provide supplementary qualitative results from our ablation study, illustrated in Fig. 4. Notably, our full model achieves superior 3D object reconstruction with coherent details.

Table 1. Comparison of performance and memory usage (GB).

Resolution	Original (256*256)			Era3D (512*512)		
Model	LGM	DG	Ours	LGM	DG	Ours
PSNR↑	17.13	17.43	**23.02**	18.06	18.51	**24.42**
SSIM↑	0.808	0.810	**0.873**	0.827	0.832	**0.886**
LPIPS↓	0.199	0.265	**0.135**	0.188	0.239	**0.112**
Mem (Training)↓	7.66	**3.14**	7.01	19.35	**7.20**	17.38
Mem (Inference)↓	6.41	**0.48**	5.62	11.83	**0.50**	10.69

Table 2. Performance with different multi-view generators.

	ImageDream		Wonder3D		
	LGM	Ours (w/o normal)	LGM	Ours (w/o normal)	Ours
PSNR↑	17.13	**19.08**	17.69	20.15	**23.02**
SSIM↑	0.808	**0.838**	0.815	0.848	**0.873**
LPIPS↓	0.199	**0.185**	0.192	0.172	**0.135**

3. Memory Usage

We report the comparisons of memory usage between our method, LGM [6] and DreamGaussian (DG) [5] (the 2nd column, the bottom 2 rows of Table 1), the top 3 methods in generation quality. Our memory usage is lower than LGM. DG has the lowest memory cost because DG is an optimization-based method for a special object without network training. Nevertheless, our method outperforms other Gaussian-based approaches by a significant margin with acceptable memory usage.

We show the comparisons of performance and memory usage on high-resolution images in the 3rd column of Table 1. Because the output resolution of most existing multi-view generators is restricted to 256×256, we adopt the latest model Era3D [3] as the multi-view generator for all methods, which can generate 512×512 resolution images. We can see our method still outperforms other Gaussian-based approaches by a significant margin with acceptable memory usage.

4. Effectiveness of Our Reconstruction Model

We compare our method with LGM [6], when using the same multi-view generator (ImageDream [7] or Wonder3D [4]), in Table 2. Our method consistently outperforms LGM on different multi-view generators, even without using normals (ImageDream does not provide normal maps). It confirms that our quality improvements over baselines like LGM are attributed to our reconstruction model rather than multi-view generators.

Table 3. User Study. The rating is from 1 to 5.

	LRM	DreamGaussian	LGM	One-2-3-45	Wonder3D	TriplaneGaussian	Ours
Image consistency↑	3.42	2.86	3.07	2.01	2.28	2.69	**4.21**
Overall quality↑	3.56	3.39	3.18	2.33	2.57	3.04	**4.01**

5. User study

We conduct a user study of 50 participants to evaluate the 3D generation quality of different methods (Table 3). For each participant, 30 groups of 360° videos rendered from objects generated by the seven models are randomly sampled from the GSO dataset. Then we ask them to rate each group's seven objects in terms of image consistency and overall quality, and calculate the mean scores. Our method is preferred among all methods.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022. 1

[2] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1

[3] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: high-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024. 1

[4] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1

[5] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1

[6] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2025. 1

[7] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 1

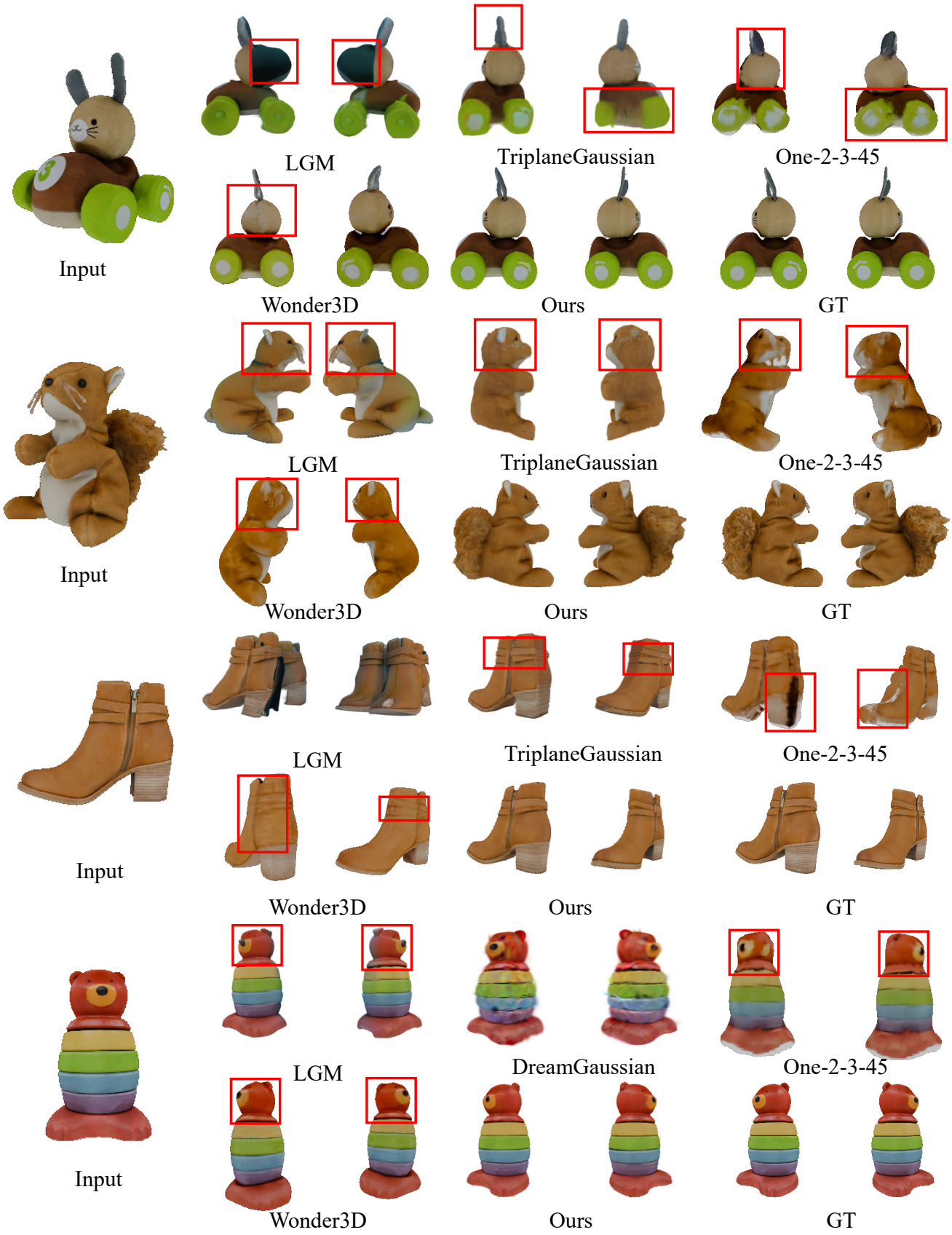

Figure 1. Qualitative comparisons of novel view synthesis between GS-RGBN and other methods on the GSO dataset.

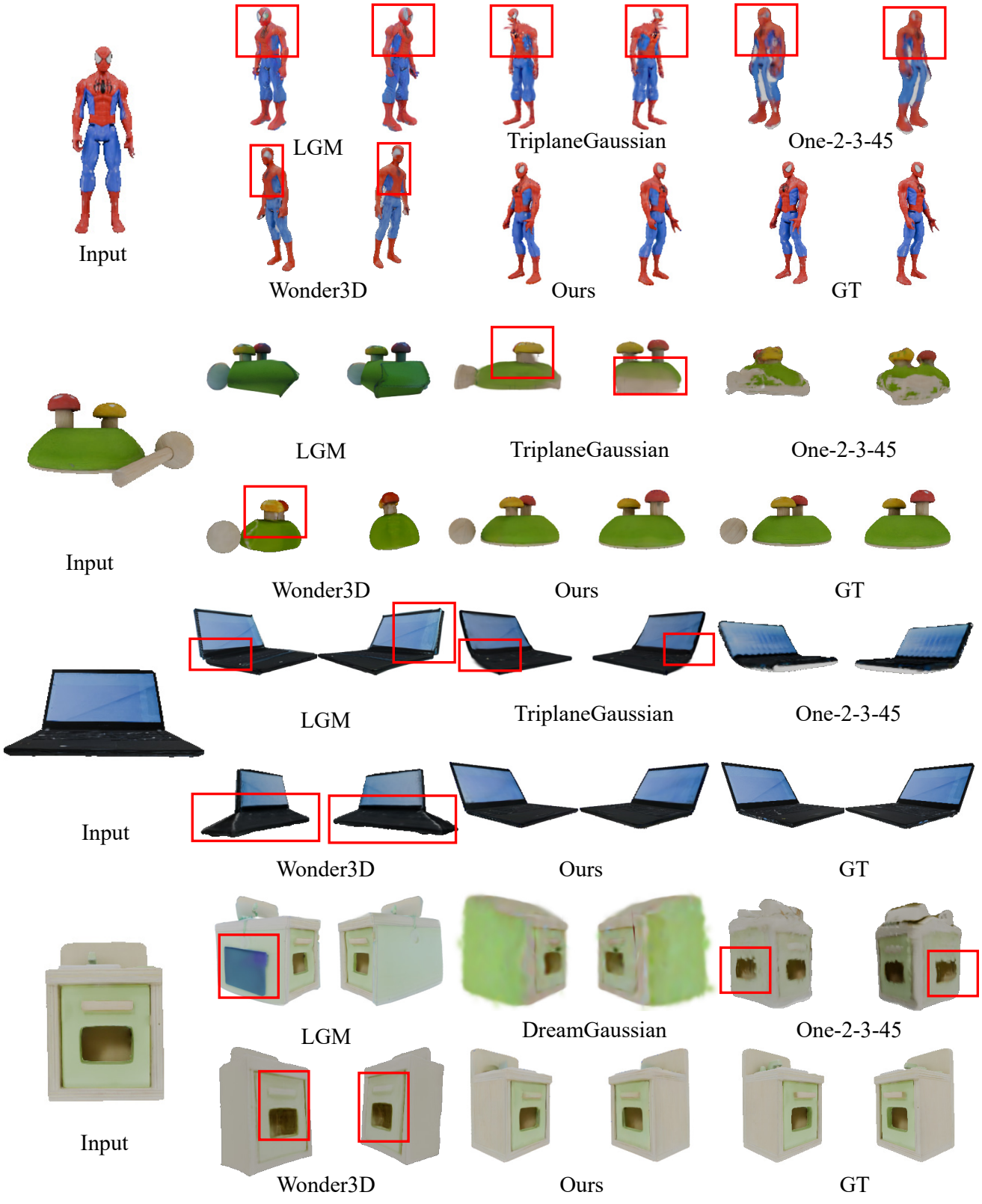

Figure 2. Qualitative comparisons of novel view synthesis between GS-RGBN and other methods on the GSO dataset.

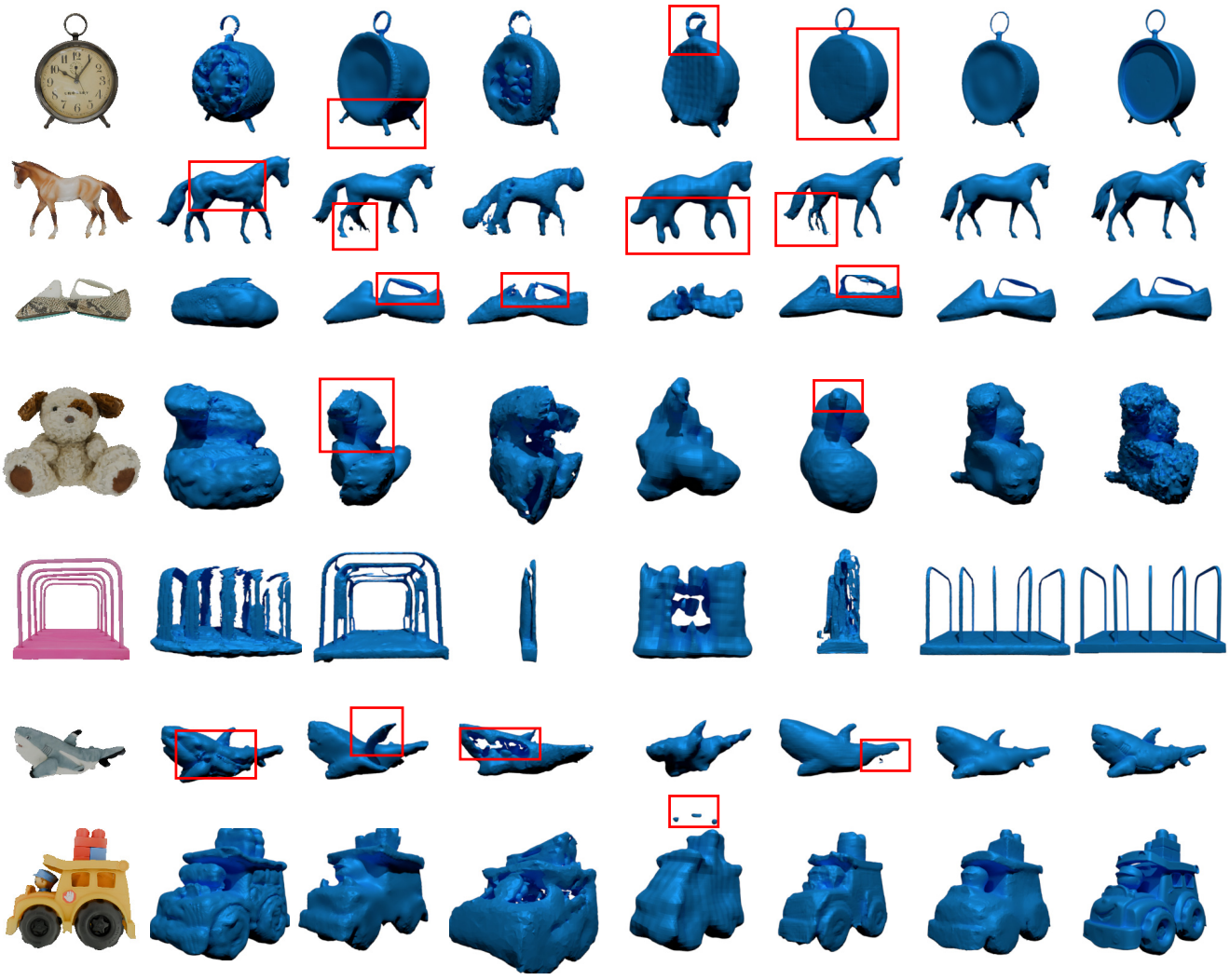

Figure 3. Qualitative comparisons of single view reconstruction between GS-RGBN and other methods on the GSO dataset.

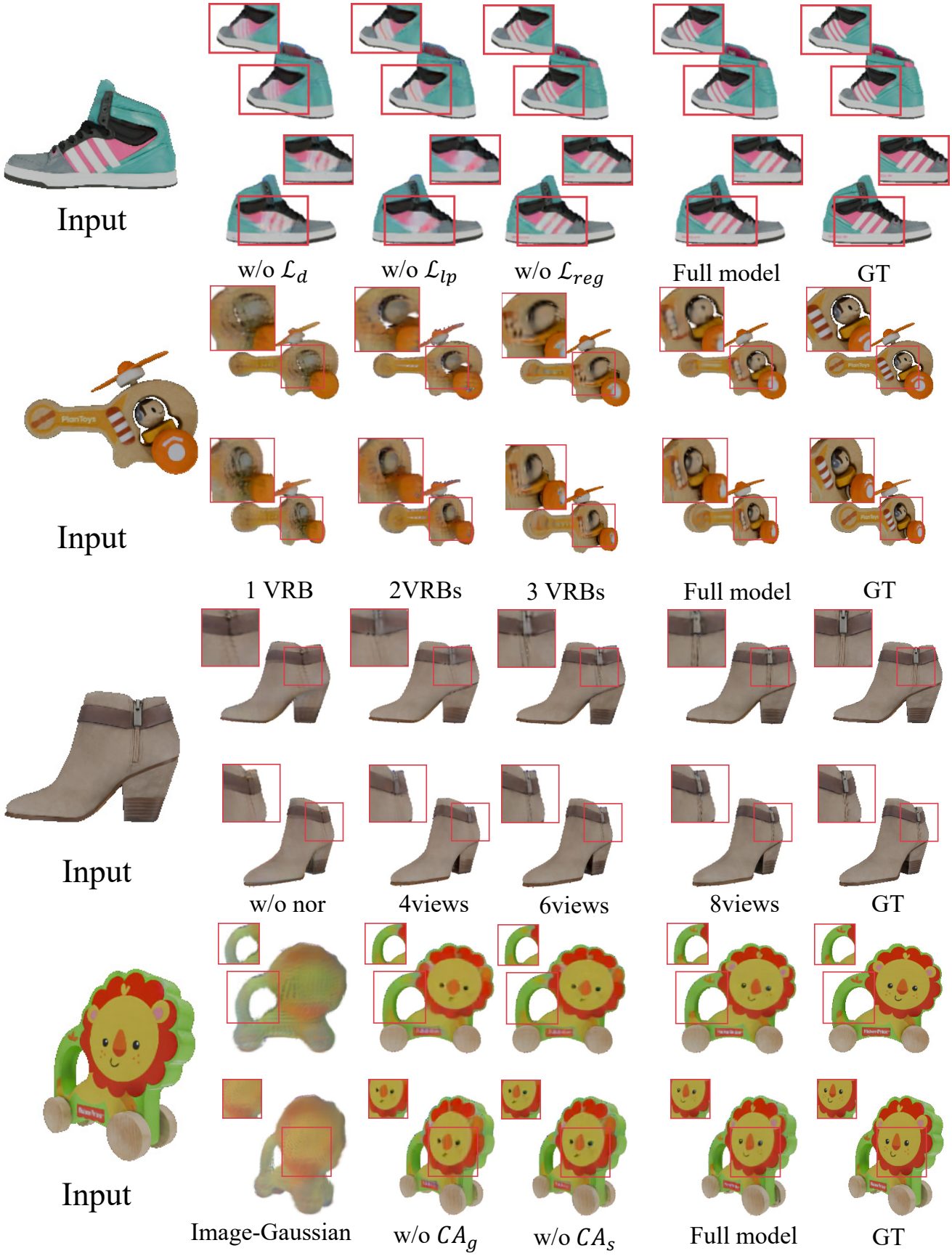

Figure 4. Ablation study of different training models. Our full model achieves the best 3D object reconstruction with consistent details.