SfM-Free 3D Gaussian Splatting via Hierarchical Training

Supplementary Material

Bo Ji Angela Yao
National University of Singapore

1. Experiments

1.1. Implementation details

For a single-image 3DGS model, we use the depth map of the input image to predict a point cloud and initialize the 3D Gaussians. The optimization process relies solely on this specific image.

When constructing a 3DGS model from a segment C_j, we set the camera pose of the first frame in C_j as the identity, i.e., $[\mathbb{I}|\mathbf{0}]$. The camera poses for subsequent frames are relative to this initial frame. We compute the monocular depth map from the first frame, predict a point cloud from this depth map, and use the resulting point cloud to initialize the 3D Gaussians. The optimization then iterates sequentially from the first to the last frame of C_j. For each new frame, we optimize the current 3DGS model using all previous frames, including the current one, with a higher probability of sampling the latest frame for training.

When merging two base 3DGS models, we prune 50% of the 3D Gaussians from each model. As a result, the number of 3D Gaussians in the merged model equals the average number of 3D Gaussians in the original base models.

During multi-source supervision, we initially train the merged 3DGS model using both the original training frames and pseudo-views generated by the two base 3DGS models. Since the quality of pseudo-views is often lower than that of interpolated images, we transition to the next training stage, where the merged 3DGS model is optimized using the original training frames and the interpolated frames.

1.2. Comparison with state-of-the-art

Reduced training time. Our proposal, especially when incorporates video frame interpolation (VFI), typically requires more training time. However, a key advantage of our proposal is that each base 3DGS model is independent. This independence enables parallel training across multiple GPUs, accelerating the process. For scenarios without distributed training, we evaluate a setting with limited training budget. Specifically, we develop a lite training version that utilizes only hierarchical training while removing VFI and supervision from pseudo frames. Additionally, we reduce the training iterations to align the training time with CF-3DGS [1]. As shown in Table 1, our approach consistently outperforms CF-3DGS across all scenes, achieving an average PSNR increase of 0.68 dB, an SSIM increase of 0.01, and a reduction in LPIPS by 0.01.

Notably, when we tested our baseline, modified from CF-3DGS, with extended training iterations, the PSNR improved by only 0.02 dB (refer to Table 5 in the main paper: Variant 3 vs. 4). This highlights the limitations of CF-3DGS due to its suboptimal 3D Gaussian distribution, which constrains its performance potential, even with additional training. In contrast, our model demonstrates the capacity for continuous improvement with increased training.

Scenes	CF-3DGS [1]			Ours (lite training)		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	30.23	0.93	0.11	30.65	0.94	0.09
Barn	31.23	0.90	0.10	31.46	0.92	0.09
Museum	29.91	0.91	0.11	30.86	0.94	0.09
Family	31.27	0.94	0.07	32.87	0.96	0.06
Horse	33.94	0.96	0.05	34.43	0.97	0.04
Ballroom	32.47	0.96	0.07	32.63	0.96	0.05
Francis	32.72	0.91	0.14	33.01	0.92	0.15
Ignatius	28.43	0.90	0.09	29.76	0.93	0.08
Mean	31.28	0.93	0.09	31.96	0.94	0.08

Table 1. **Novel view synthesis results on Tanks and Temples [2] .** Adjusting the training time to match that of CF-3DGS, our approach consistently delivers superior performance.

Qualitative comparison. More visual comparisons are shown in Fig. 1 to 7. Our proposal consistently achieves the superior performance and reduces artifacts.

1.3. Ablation study

In this supplementary material, we present per-scene results corresponding to the ablation study in Table 5 of the main paper, along with additional ablation study results. All experiments are conducted on the Tanks and Temples dataset [2].

Prune ratio. Table 2 presents the results for different pruning ratios applied to each base 3DGS model before merging. In general, higher pruning ratios result in a more compact representation with reduced memory storage but also lead to a slight performance drop. Interestingly, the best performance is not achieved with a pruning ratio of 0% (no pruning). At 0% pruning, the memory usage (1.34GB)

Scenes	Prune 0%				Prune 25%				Prune 50%				Prune 75%			
	PSNR ↑	SSIM ↑	LPIPS ↓	Mem ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Mem ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Mem ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Mem ↓
Church	31.27	0.94	0.08	1.37	31.49	0.94	0.08	0.87	31.67	0.95	0.08	0.85	31.51	0.94	0.08	0.80
Barn	32.18	0.92	0.08	1.66	32.15	0.92	0.08	1.48	32.27	0.92	0.08	1.41	32.20	0.92	0.08	1.37
Museum	31.99	0.95	0.07	1.29	31.83	0.94	0.07	1.20	31.75	0.94	0.07	1.10	31.56	0.94	0.08	1.04
Family	34.14	0.97	0.05	1.38	34.29	0.97	0.05	1.23	34.20	0.97	0.05	1.21	33.91	0.97	0.05	1.13
Horse	35.65	0.98	0.04	1.09	35.76	0.98	0.03	0.98	35.44	0.98	0.04	0.90	35.62	0.98	0.04	0.85
Ballroom	33.66	0.97	0.04	1.33	33.57	0.97	0.04	1.22	33.41	0.97	0.04	1.14	33.36	0.97	0.04	1.09
Francis	33.63	0.92	0.13	0.80	33.69	0.92	0.13	0.67	33.66	0.92	0.13	0.75	33.60	0.92	0.13	0.64
Ignatius	31.43	0.94	0.06	1.78	31.59	0.94	0.06	1.50	31.78	0.94	0.06	1.43	31.46	0.94	0.06	1.38
Mean	32.99	0.95	0.07	1.34	33.05	0.95	0.07	1.14	33.02	0.95	0.07	1.10	32.90	0.95	0.07	1.04

Table 2. **Ablation study of the pruning ratio.** The best performance is achieved at a 25% pruning ratio. Generally, pruning more 3D Gaussians leads to a slight performance drop but provides a more compact representation with significantly reduced memory usage.

is significantly higher compared to other pruning levels (1.04–1.14GB). We hypothesize that without pruning, a large number of unimportant 3D Gaussians, which do not contribute much to the original base 3DGS model, remain in the model and continue to hinder the optimization of the merged model. By removing these redundant Gaussians, our pruning strategy allows the optimization to focus on the more critical 3D Gaussians, thereby enhancing overall performance.

CF-3DGS vs. CF-3DGS + VFI. We compare CF-3DGS [1] with CF-3DGS enhanced by VFI [3]. The addition of VFI improves the average PSNR by 0.17 dB, with the most significant gain observed in the *Barn* scene. This is because the *Barn* scene has a relatively low frame rate and large camera motion, making the interpolated frames crucial for performance improvement.

Scenes	CF-3DGS [1]			CF-3DGS + VFI [3]		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	30.23	0.93	0.11	30.01	0.93	0.10
Barn	31.23	0.90	0.10	33.51	0.94	0.07
Museum	29.91	0.91	0.11	28.87	0.89	0.12
Family	31.27	0.94	0.07	32.44	0.96	0.06
Horse	33.94	0.96	0.05	33.66	0.96	0.06
Ballroom	32.47	0.96	0.07	32.03	0.96	0.05
Francis	32.72	0.91	0.14	32.75	0.92	0.14
Ignatius	28.43	0.90	0.09	28.29	0.91	0.10
Mean	31.28	0.93	0.09	31.45	0.93	0.09

Table 3. **Ablation study of video frame interpolation on CF-3DGS.** VFI improves the average PSNR by 0.17dB.

Global training. Table 4 presents an ablation study comparing the performance of baseline and global training. The results show that global training provides marginal improvements. Notable gains are observed in metrics like PSNR for scenes such as *Church* and *Barn*, and a slight reduction in LPIPS for *Ballroom*.

Progressive vs. hierarchical training. Table 5 compares the performance of progressive and hierarchical training.

Scenes	Baseline			Global training		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	30.44	0.93	0.09	31.06	0.94	0.08
Barn	30.09	0.88	0.11	31.22	0.90	0.10
Museum	30.24	0.91	0.10	29.37	0.90	0.10
Family	33.12	0.96	0.05	33.29	0.96	0.05
Horse	34.08	0.96	0.05	34.34	0.97	0.05
Ballroom	32.82	0.96	0.05	32.89	0.96	0.04
Francis	32.84	0.92	0.14	32.66	0.91	0.14
Ignatius	28.37	0.91	0.09	27.30	0.88	0.10
Mean	31.50	0.93	0.09	31.52	0.93	0.08

Table 4. **Ablation study of global training.** Increasing the training time only yields a marginal improvement.

The results indicate that hierarchical training improves the overall mean performance.

Scenes	Progressive training			Hierarchical training		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	31.81	0.95	0.08	31.67	0.95	0.08
Barn	34.06	0.96	0.05	32.27	0.92	0.08
Museum	30.74	0.94	0.08	31.75	0.94	0.07
Family	34.36	0.97	0.05	34.20	0.97	0.05
Horse	34.66	0.98	0.04	35.44	0.98	0.04
Ballroom	33.75	0.97	0.04	33.41	0.97	0.04
Francis	34.07	0.93	0.13	33.66	0.92	0.13
Ignatius	29.14	0.93	0.08	31.78	0.94	0.06
Mean	32.82	0.95	0.07	33.02	0.95	0.07

Table 5. **Ablation study of progressive and hierarchical training.** Hierarchical training outperforms the progressive training.

Video frame interpolation. Table 6 evaluates the impact of video frame interpolation (VFI) on the 'Baseline + Hierarchical Training (HT)' approach. Adding VFI slightly improves the mean PSNR (33.02 to 33.37) while maintaining similar SSIM and LPIPS.

Supervision from base 3DGS models. Table 7 compares the performance of 'Baseline + HT + VFI' with the proposed method that incorporates supervision from base

Scenes	Baseline + HT			Baseline + HT + VFI		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	31.67	0.95	0.08	31.45	0.94	0.08
Barn	32.27	0.92	0.08	34.47	0.96	0.05
Museum	31.75	0.94	0.07	31.65	0.94	0.08
Family	34.20	0.97	0.05	34.02	0.97	0.05
Horse	35.44	0.98	0.04	35.90	0.98	0.04
Ballroom	33.41	0.97	0.04	33.71	0.97	0.04
Francis	33.66	0.92	0.13	34.02	0.93	0.13
Ignatius	31.78	0.94	0.06	31.71	0.94	0.06
Mean	33.02	0.95	0.07	33.37	0.95	0.07

Table 6. **Ablation study of video frame interpolation.** VFI achieves superior performance.

3DGS models. The proposed method achieves higher mean PSNR (33.53 vs. 33.37) and SSIM (0.96 vs. 0.95) while maintaining the same LPIPS value (0.07).

Scenes	Baseline + HT + VFI			Ours		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	31.45	0.94	0.08	31.34	0.94	0.08
Barn	34.47	0.96	0.05	34.95	0.97	0.05
Museum	31.65	0.94	0.08	31.59	0.95	0.08
Family	34.02	0.97	0.05	34.71	0.97	0.05
Horse	35.90	0.98	0.04	35.82	0.98	0.03
Ballroom	33.71	0.97	0.04	34.12	0.97	0.04
Francis	34.02	0.93	0.13	34.09	0.93	0.13
Ignatius	31.71	0.94	0.06	31.64	0.95	0.06
Mean	33.37	0.95	0.07	33.53	0.96	0.07

Table 7. **Ablation study of supervision from base 3DGS models.** Incorporating the supervision from base 3DGS models yields better performance.

Unknown camera intrinsics. We experiment with heuristics instead of known camera intrinsics by setting the FoV to 70°. Table 8 compares performance with and without incorporating camera intrinsics. The results show that including camera intrinsics significantly improves mean PSNR (33.53 vs. 32.17), SSIM (0.96 vs. 0.94), and LPIPS (0.07 vs. 0.09). Inaccurate camera intrinsics hinder pose estimation and may introduce scale ambiguity.

Scenes	w/o intrinsic			Ours		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Church	30.53	0.94	0.10	31.34	0.94	0.08
Barn	33.22	0.95	0.07	34.95	0.97	0.05
Museum	30.50	0.92	0.11	31.59	0.95	0.08
Family	32.64	0.95	0.08	34.71	0.97	0.05
Horse	35.71	0.98	0.04	35.82	0.98	0.03
Ballroom	33.56	0.97	0.04	34.12	0.97	0.04
Francis	32.71	0.91	0.15	34.09	0.93	0.13
Ignatius	28.46	0.91	0.11	31.64	0.95	0.06
Mean	32.17	0.94	0.09	33.53	0.96	0.07

Table 8. **Ablation study of camera intrinsics.** Known camera intrinsics positively impact the performance.

References

[1] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 1, 2

[2] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 2017. 1, 4, 5, 6, 7

[3] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 2

[4] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 8, 9, 10

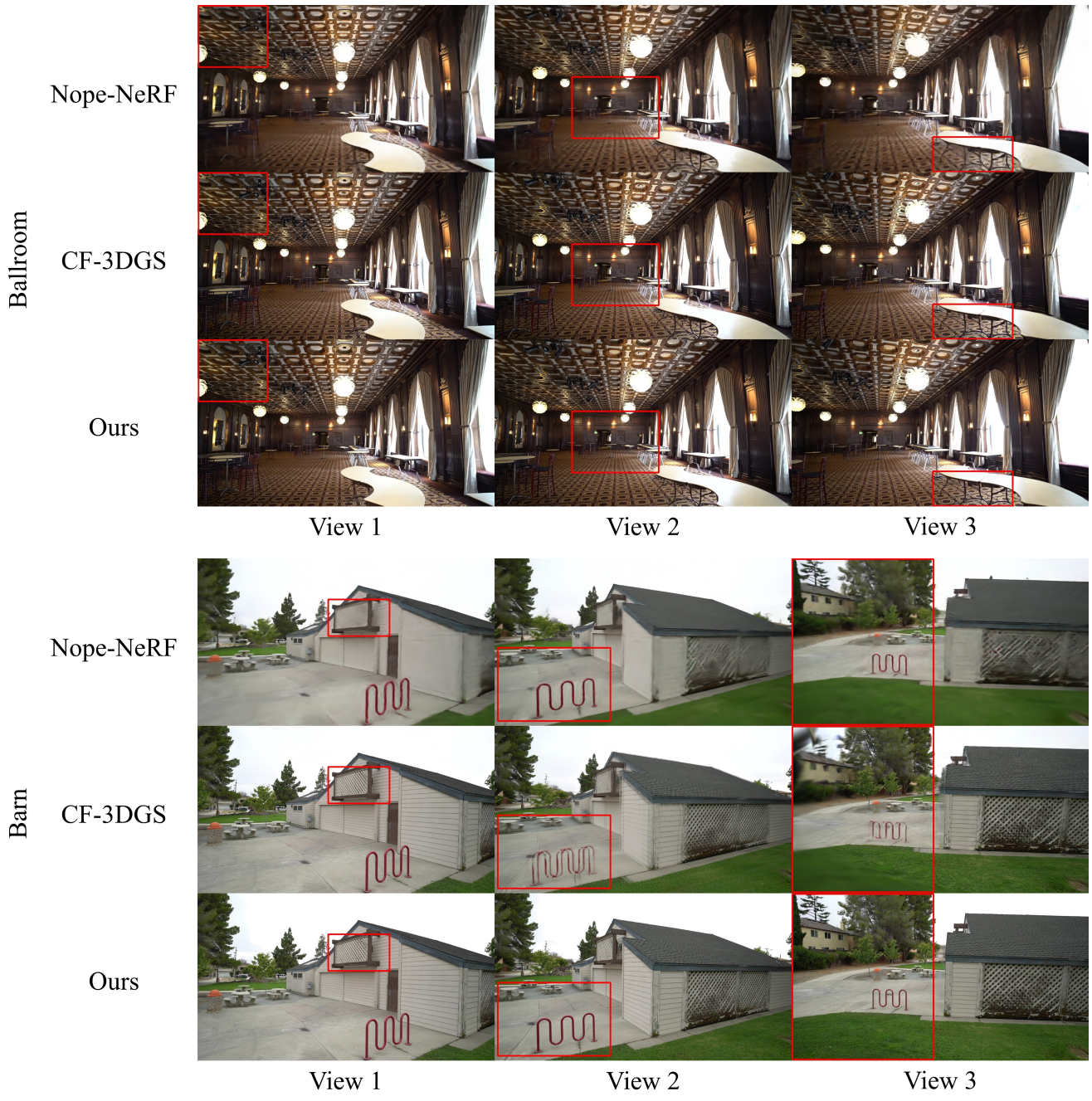

Figure 1. **Qualitative novel view synthesis results on Tanks and Temples [2].** Please zoom in for a better view.

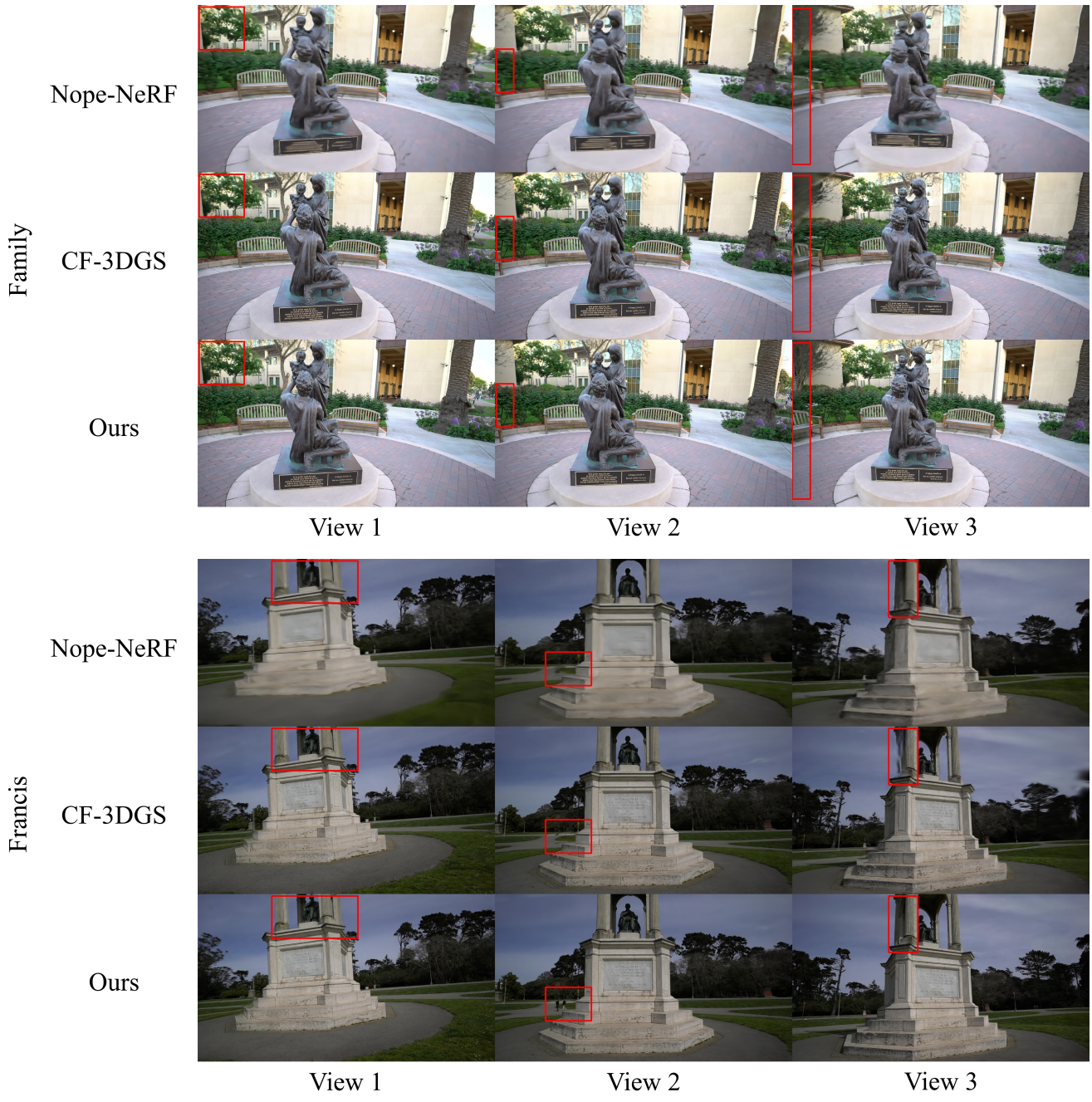

Figure 2. **Qualitative novel view synthesis results on Tanks and Temples [2].** Please zoom in for a better view.

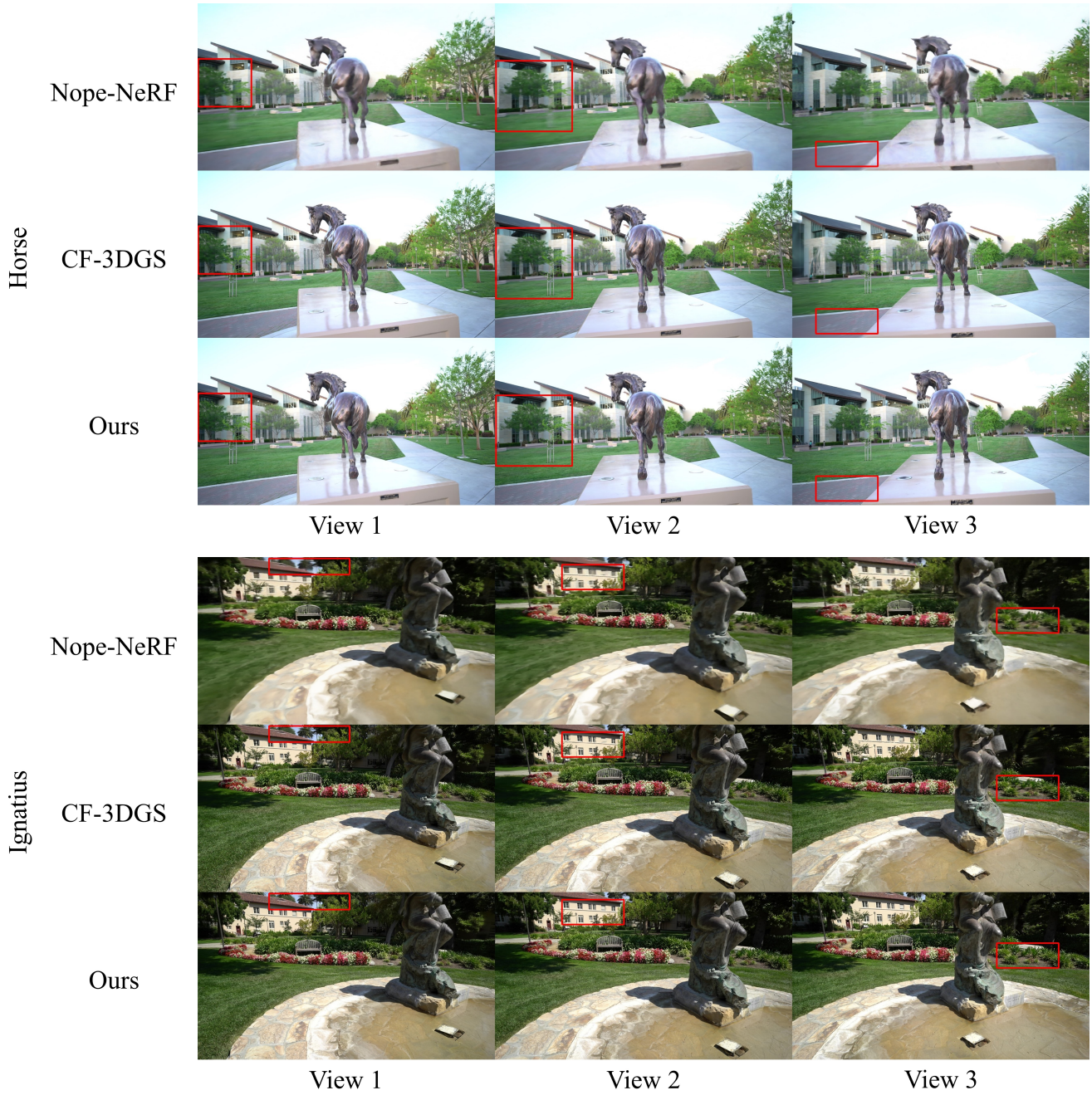

Figure 3. **Qualitative novel view synthesis results on Tanks and Temples [2].** Please zoom in for a better view.

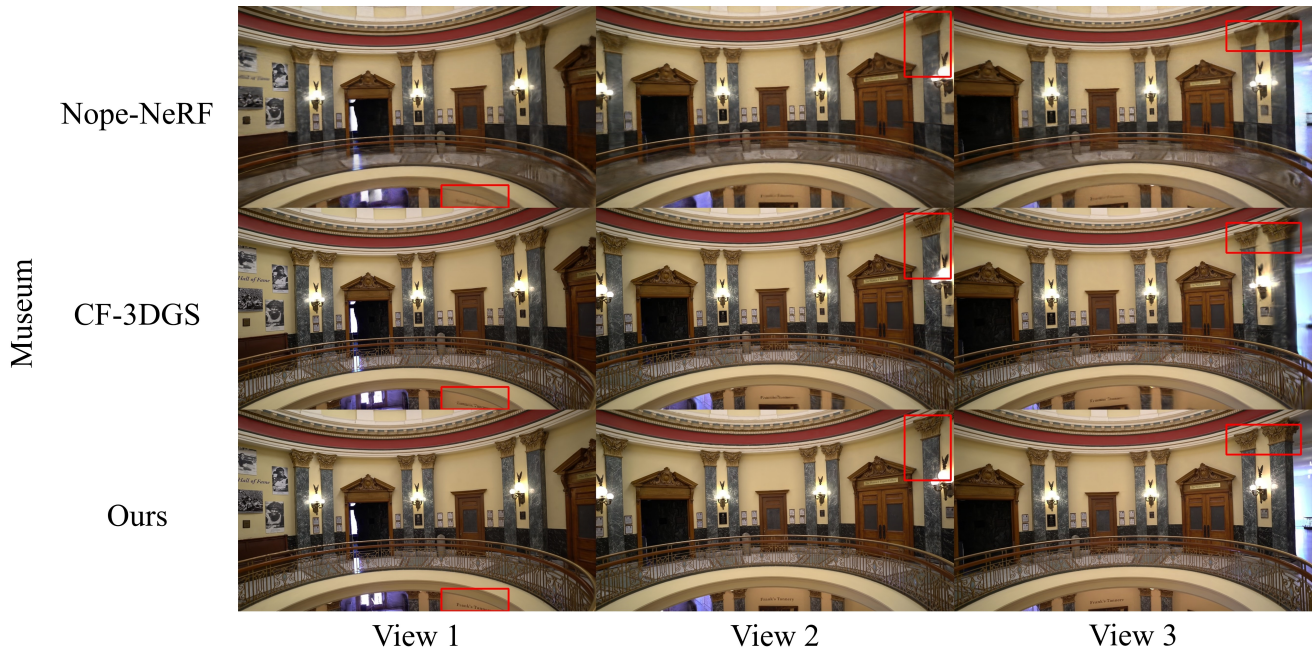

Figure 4. **Qualitative novel view synthesis results on Tanks and Temples [2].** Please zoom in for a better view.

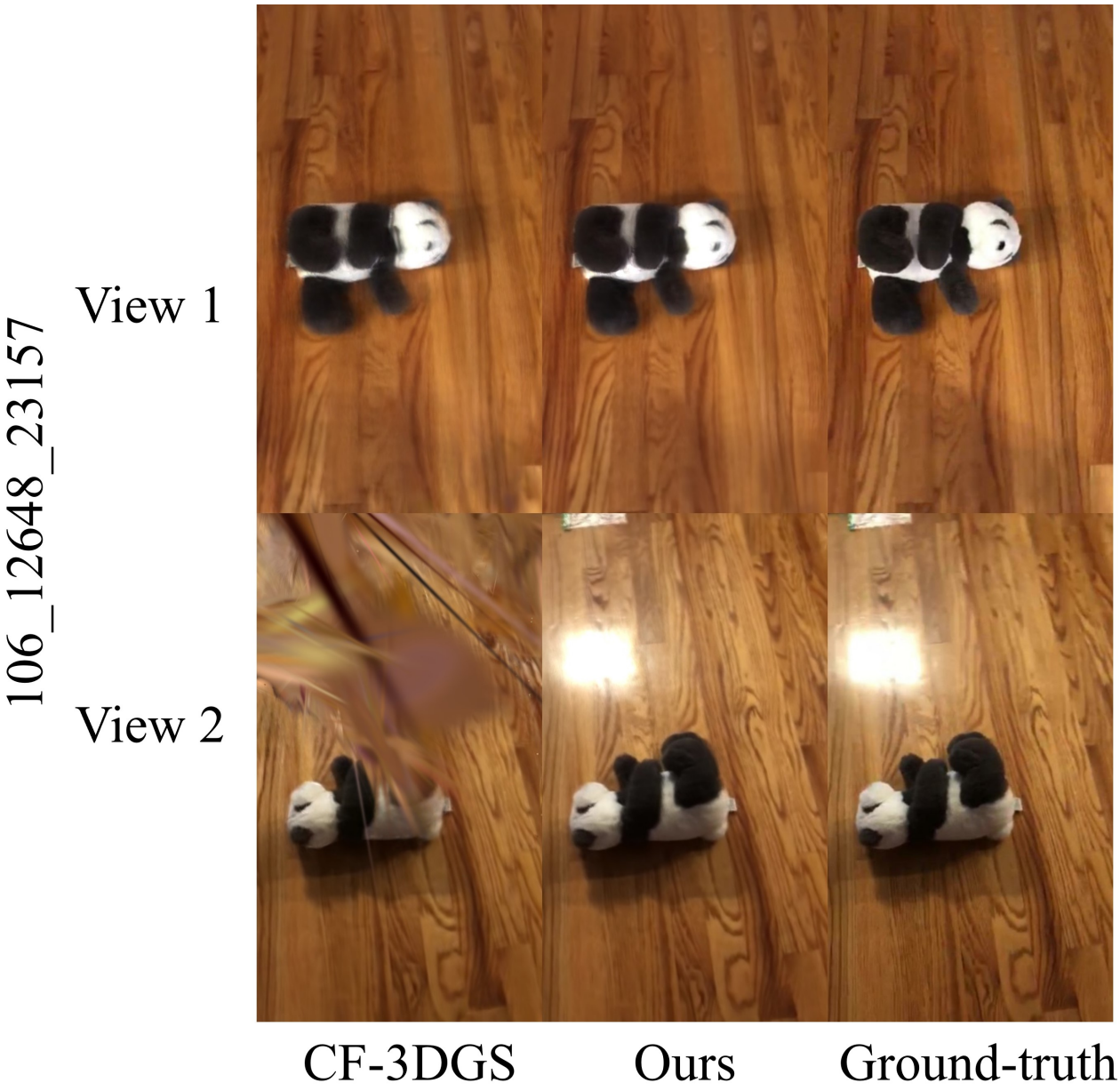

Figure 5. **Qualitative novel view synthesis results on CO3D-V2 [4].** Please zoom in for a better view.

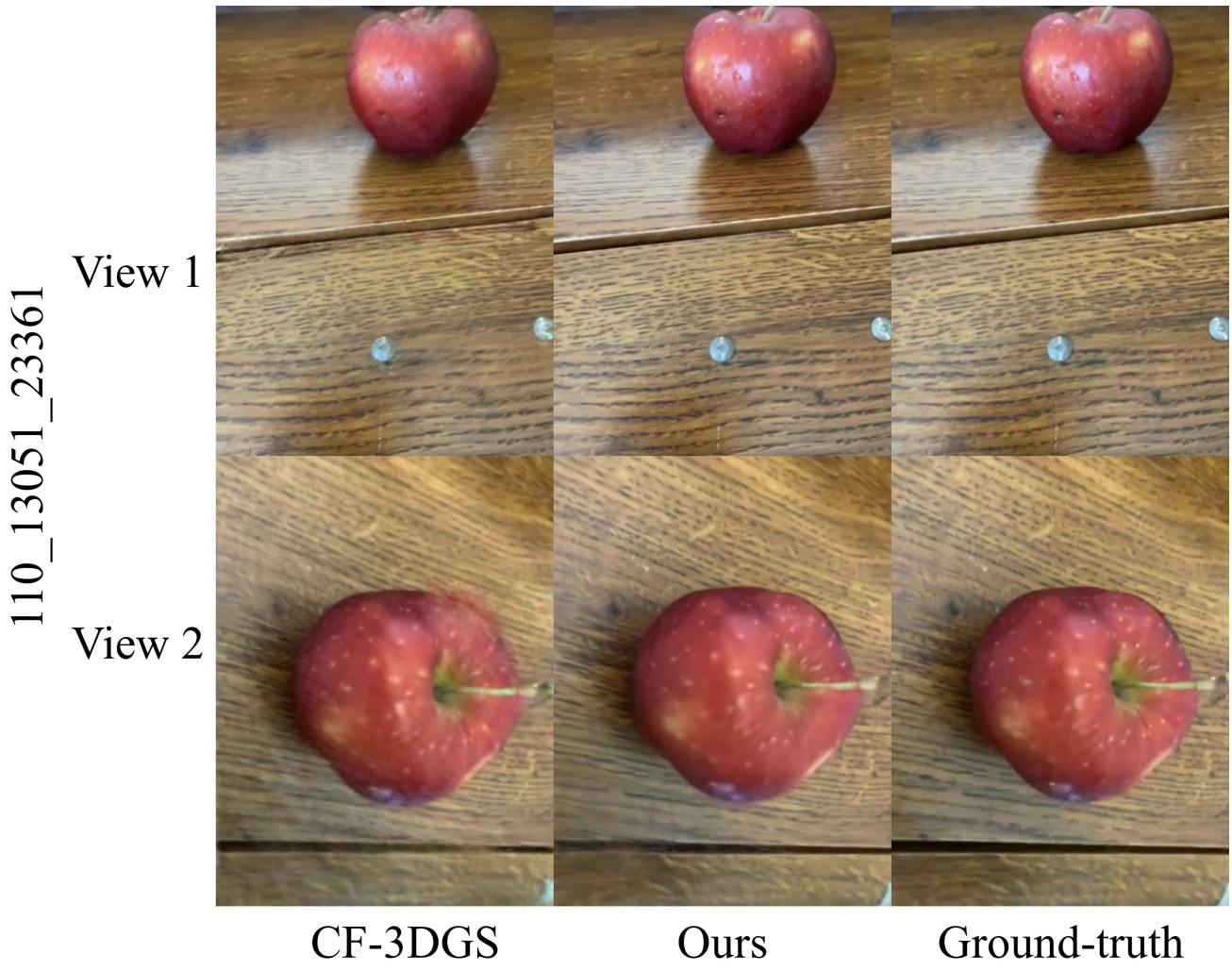

Figure 6. **Qualitative novel view synthesis results on CO3D-V2 [4].** Please zoom in for a better view.

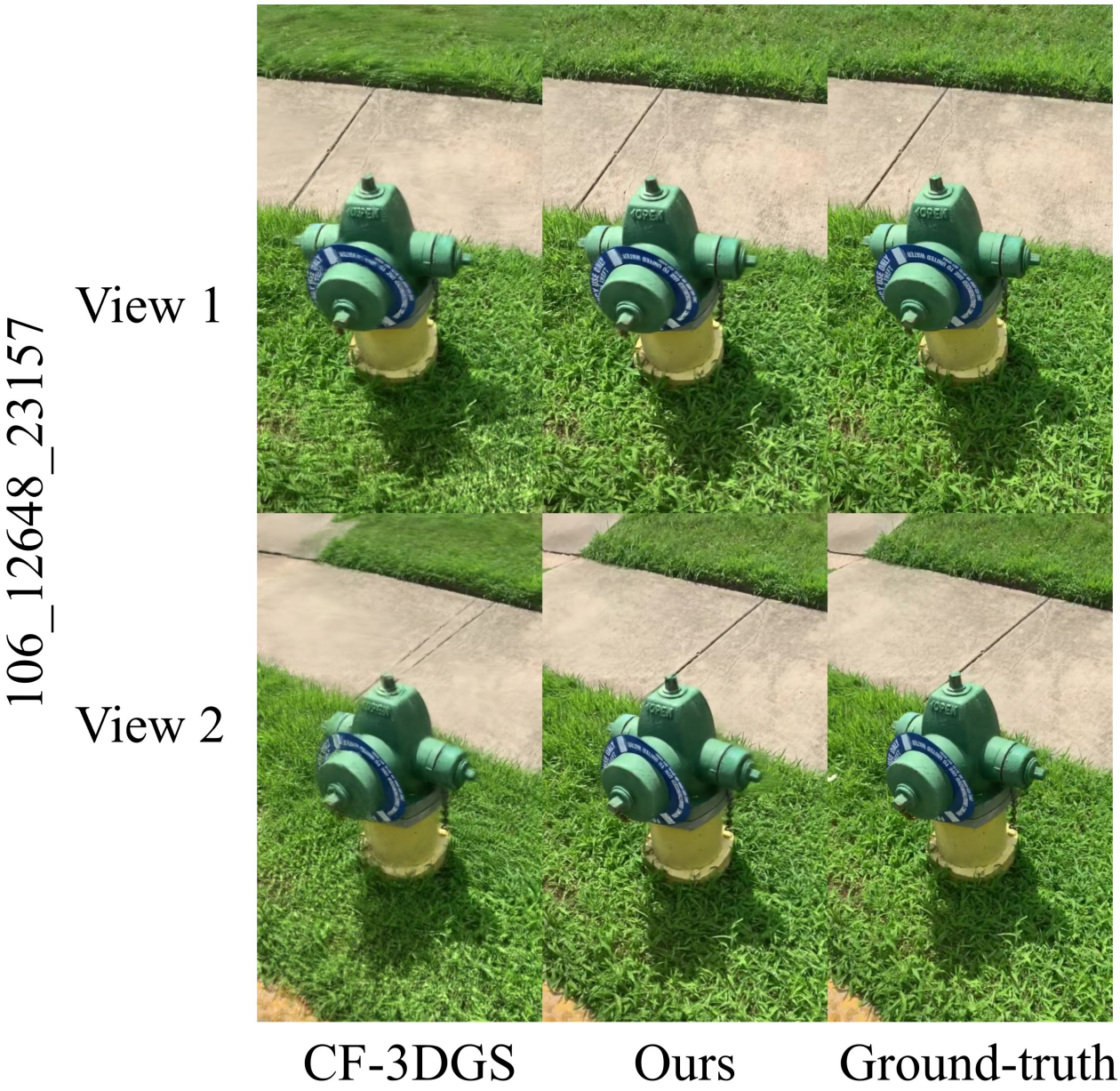

Figure 7. **Qualitative novel view synthesis results on CO3D-V2 [4].** Please zoom in for a better view.