COLMAP-Free 3D Gaussian Splatting

Supplementary Material

A. Implementation Details

The following sections include more details about the datasets we use, our training and evaluation procedure.

A.1. Dataset

We select sequences containing dramatic camera motions Tanks and Tamples [3] and CO3D-V2 [5] for training and evaluation. The details of each sequence are listed in Table 1, where *Max rotation* denotes the maximum relative rotation angle between any two frames in a sequence. The sampled images are further split into training and test sets. Starting from the 5*th* image, we sample every 8*th* image in a sequence as a test image. However, this leads to a change in the sampling rate in the temporal domain among training images. In order to study the effect of the sampling rate changes, we follow the experiment setting proposed by [2]. Specifically, for scene *Family* in Tanks and Temples [3], we sample every other image as test images, *i.e.*, training on images with odd frame ids and testing on images with even frame ids. For CO3D-V2 [5], we randomly select 10 scenes from 6 categories, *e.g.*, apple, bench, hydrant, plant, skateboard, and teddybear. The selected sequence IDs are also shown in Table 1 (bottom part). Compared to Tanks and Temples, most scenes achieve the *Max rotation* of 180° indicating more dramatic and larger camera motions than Tanks and Temples.

	Scenes	Type	Seq. length	Frame rate	Max. rotation (deg)
Tanks and Temples	Church	indoor	400	30	37.3
	Barn	outdoor	150	10	47.5
	Museum	indoor	100	10	76.2
	Family	outdoor	200	30	35.4
	Horse	outdoor	120	20	39.0
	Ballroom	indoor	150	20	30.3
	Francis	outdoor	150	10	47.5
	Ignatius	outdoor	120	20	26.0
CO3D-V2	34_1403_4393	indoor	202	30	180.0
	106_12648_23157	outdoor	202	30	180.0
	110_13051_23361	indoor	202	30	71.6
	219_23121_48537	indoor	202	30	180.0
	245_26182_52130	indoor	202	30	180.0
	247_26441_50907	indoor	202	30	180.0
	407_54965_106262	indoor	202	30	180.0
	415_57112_110099	outdoor	202	30	180.0
	415_57121_110109	outdoor	202	30	180.0
	429_60388_117059	outdoor	202	30	180.0

Table 1. **Details of selected sequences.** We downsample several videos to a lower frame rate. FPS denotes frame per second. *Max rotation* denotes the maximum relative rotation angle between any two frames in a sequence. Our method can handle dramatic camera motion (large maximum rotation angle).

Algorithm 1 Local 3DGS Optimization

$\{I_t, I_{t+1}\} \leftarrow$ Two nearby images
DPT $\leftarrow$ Monocular Depth Estimation Model
$D_t \leftarrow \text{DPT}(I_t)$
$G_t \leftarrow \text{InitGauss}(I_t, D_t)$ ▷ Init Local 3DGS
$T_t \leftarrow$ Identity $\mathbb{I}$ ▷ Init Pose
while not converged **do**
  $\hat{I}_t \leftarrow \text{Rasterize}(G_t)$
  $L \leftarrow \text{Loss}(I_t, \hat{I}_t)$
  $G_t \leftarrow \text{Adam}(\nabla L)$ ▷ Update Local 3DGS
end while
while not converged **do**
  $\hat{I}_{t+1} \leftarrow \text{Rasterize}(T_t \odot G_t)$
  $L \leftarrow \text{Loss}(I_{t+1}, \hat{I}_{t+1})$
  $T_t^* \leftarrow \text{Adam}(\nabla L)$ ▷ Update Pose
end while
$T_t \leftarrow \prod_{i=1}^{t} T_i$ ▷ Output Pose

A.2. Training Details.

Local 3DGS. During the training of local 3DGS, we first obtain the monocular depth map of the input image by pre-trained monocular depth estimator, *i.e.*, DPT [4], ZeoDepth [1]. Then, the depth map is lifted up with the given camera intrinsic. As the high-resolution input images could lead to a huge amount of point clouds, we downsample the point cloud first before fitting it by 3DGS. Then, the downsampled point cloud is used to initialize the local 3DGS and is further optimized on the input view via photometric loss for 500 iterations. To obtain the transformation of the 3D Gaussian between two views, we freeze the pre-trained local 3DGS including all attributes (*i.e.*, position, SH coefficient, opacity, scale, and rotation), and learn the pose parameter of a quaternion vector a translation vector by the photometric loss between the target view and the rendering image. In detail, the freeze local 3D Gaussian is first transformed into the target view coordinate by the learnable pose parameter and then rendered into the target view by the gaussian splatting. The optimization of the camera pose learning process takes 300 steps. The optimization algorithm of local 3DGS is summarized in Algorithm 1

Global 3DGS. The optimization process of the global 3DGS is comprehensively detailed in Algorithm 2. Specifically, it starts and initializes from the first frame and its monocular depth estimation. Subsequently, camera poses are estimated in a sequential manner using the local 3DGS, as described in Algorithm 1. Concurrently, the global 3DGS is updated with all the observed images to date (*i.e.*, from the first to the cur-

Algorithm 2 COLMAP-Free 3DGS Optimization

$\{I_t | t = 1...N\} \leftarrow$ Image sequence
DPT $\leftarrow$ Monocular Depth Estimation Model
$D_1 \leftarrow \text{DPT}(I_1)$
$G \leftarrow \text{InitGauss}(I_1, D_1)$ ▷ Init Globla Gauss
$i \leftarrow 0$ ▷ Iteration Count
for all Image I_t **in** $I_{t=1...N}$ **do**
  $T_t \leftarrow$ Local 3DGS(I_t, I_{t+1}) ▷ Eestimate Pose
  while not converged **do**
    $T_j, I_j \leftarrow \text{SampleTrainingView}()$ ▷ $j \leq t$
    $\hat{I}_j \leftarrow \text{Rasterize}(G, T_j)$
    $L \leftarrow Loss(I_j, \hat{I}_j)$
    $G \leftarrow \text{Adam}(\nabla L)$ ▷ Update Gauss
    $i \leftarrow i + 1$
  end while
  for all Gaussians (μ, Σ, c, α) **in** G **do**
    if $\nabla_p L > \tau_p$ **then** ▷ Densification
      SplitGaussian(μ, Σ, c, α)
      CloneGaussian(μ, Σ, c, α)
    end if
    if $\alpha < \epsilon$ or IsTooLarge(μ, Σ) **then** ▷ Pruning
      RemoveGaussian()
    end if
  end for
end for

rent image), in tandem with the camera pose estimation. As each new frame is introduced, the global 3DGS progressively grows and expands through a densification process.

A.3. Evaluation Metrics

Novel View Synthesis. We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [7], and Learned Perceptual Image Patch Similarity (LPIPS) [8] to measure the novel view synthesis quality. For LPIPS, we use a VGG architecture [6].

Pose Accuracy. To evaluate pose accuracy, we employ standard visual odometry metrics, including Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). ATE quantifies the discrepancy between estimated camera positions and their ground truth counterparts. RPE, on the other hand, assesses the errors in relative poses between image pairs. This includes both relative rotation error (RPE_r) and relative translation error (RPE_t).

B. Additional Experiments

The subsequent sections present further quantitative and qualitative results of novel view synthesis and camera pose estimation, conducted on both the Tanks and Temples and CO3D-V2 datasets.

scenes	Nope-NeRF			Ours		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
189_20393_38136	29.37	0.85	0.54	32.41	0.92	0.26
247_26441_50907	23.49	0.73	0.54	23.88	0.75	0.36
407_54965_106262	25.53	0.83	0.58	27.80	0.84	0.35
429_60388_117059	22.19	0.62	0.56	24.44	0.68	0.36
46_2587_7531	25.3	0.73	0.46	25.44	0.80	0.21
mean	25.18	0.75	0.54	**26.79**	**0.80**	**0.31**

Table 2. **Novel view synthesis results on CO3D V2.** The best results are highlighted in bold.

B.1. Novel View Synthesis.

Render Novel Views. As mentioned in the main paper, we minimize the photometric error of the synthesized images while freezing the 3DGS model to obtain the testing camera poses. Because the test views are sampled from videos that are close to the training views, these good results may be obtained due to overfitting to the training images. Therefore, we conduct an additional qualitative evaluation on more novel views. Specifically, we fit a bezier curve from the estimated training poses and sample interpolated poses for each method to render novel view videos. The rendered images are shown in Fig. 1 and Fig. 2. Compared to Nope-NeRF [2], our approach renders photo-realistic images with more details (please check the highlighted regions).

Unknown camera intrinsic. We also conduct experiments with heuristic camera intrinsic, where we set the FoV of all scenes to 79° and make the principle points to the image center. The quantitative results are listed in the following table. We find that by setting the camera intrinsic heuristically, the performance on novel view synthesis (NVS) and camera pose estimation slightly degenerates which is reasonable as the intrinsic parameters are also important and could be further optimized along with the camera extrinsic parameters.

Method	PSNR	SSIM	LPIPS	RPE_t	RPE_r	ATE
Heuristic Intrinsic	30.90	0.92	0.09	0.044	0.072	0.004
G.T. Intrinsic	**31.28**	**0.93**	**0.09**	**0.041**	**0.069**	**0.004**

Different monocular depth estimator. We conduct ablation studies on different monocular depth estimation algorithms in the following table. We notice that more accurate monocular depth estimation results could always lead to better performance.

scenes	ZeoDepth				DepthAnything			
	PSNR	SSIM	RPE_t	RPE_r	PSNR	SSIM	RPE_t	RPE_r
Church	30.49	0.93	0.012	0.033	30.66	0.93	0.012	0.029
Barn	28.34	0.86	0.039	0.057	30.54	0.88	0.034	0.113
Museum	30.40	0.91	0.052	0.158	30.92	0.92	0.043	0.130
Family	28.79	0.91	0.093	0.037	32.54	0.95	0.037	0.069
Horse	33.32	0.95	0.101	0.035	33.96	0.96	0.108	0.075
Ballroom	32.86	0.96	0.021	0.032	32.54	0.96	0.022	0.030
Francis	31.05	0.89	0.057	0.086	32.73	0.91	0.027	0.126
Ignatius	22.75	0.75	0.172	0.083	28.89	0.89	0.043	0.075
mean	29.75	0.90	0.068	0.065	**31.60**	**0.93**	**0.041**	0.081

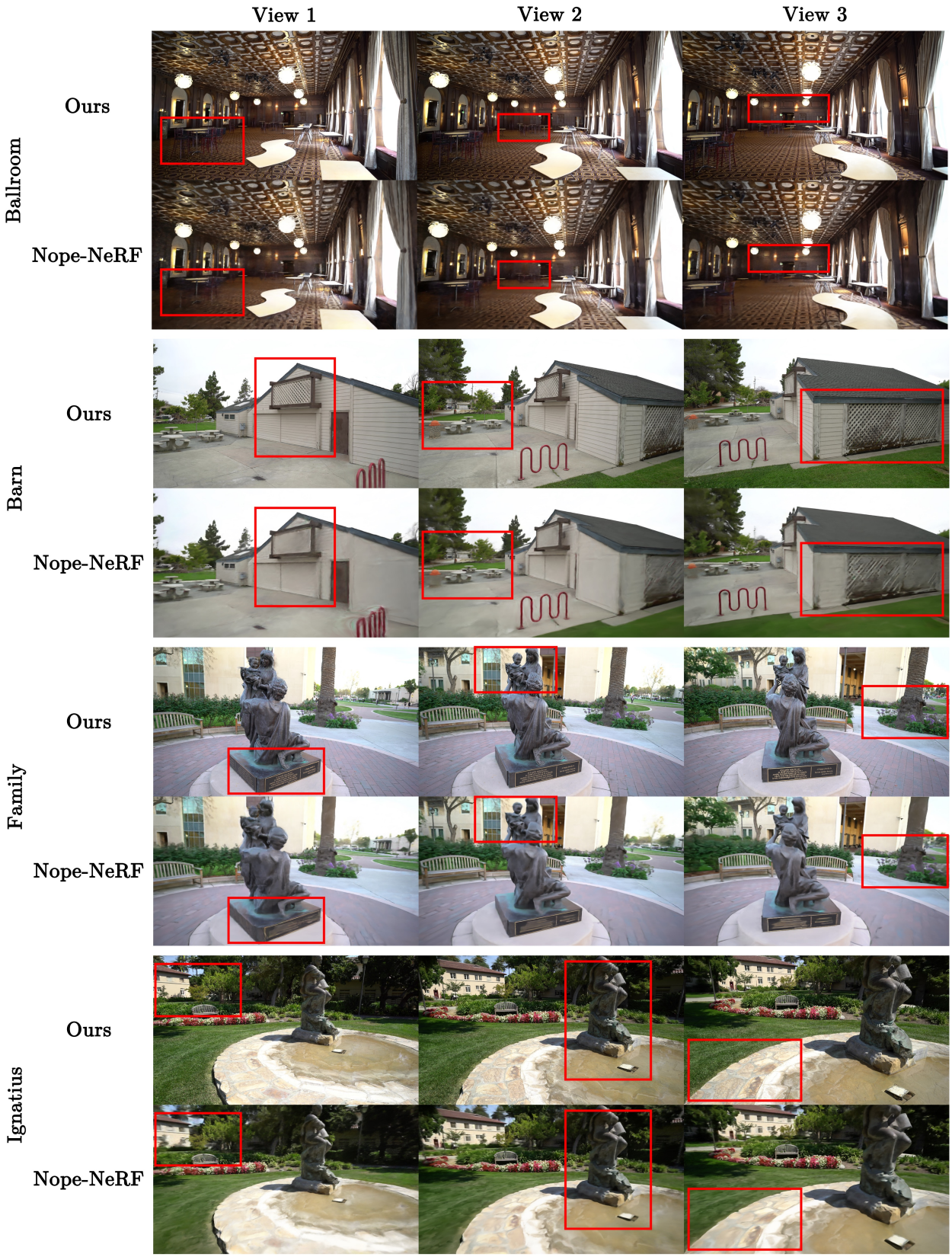

Figure 1. **Qualitative comparison for novel view synthesis on Tanks and Temples.** For each method, we fit the learned trajectory with a bezier curve and uniformly sample new viewpoints for rendering. **Better viewed when zoomed in.**

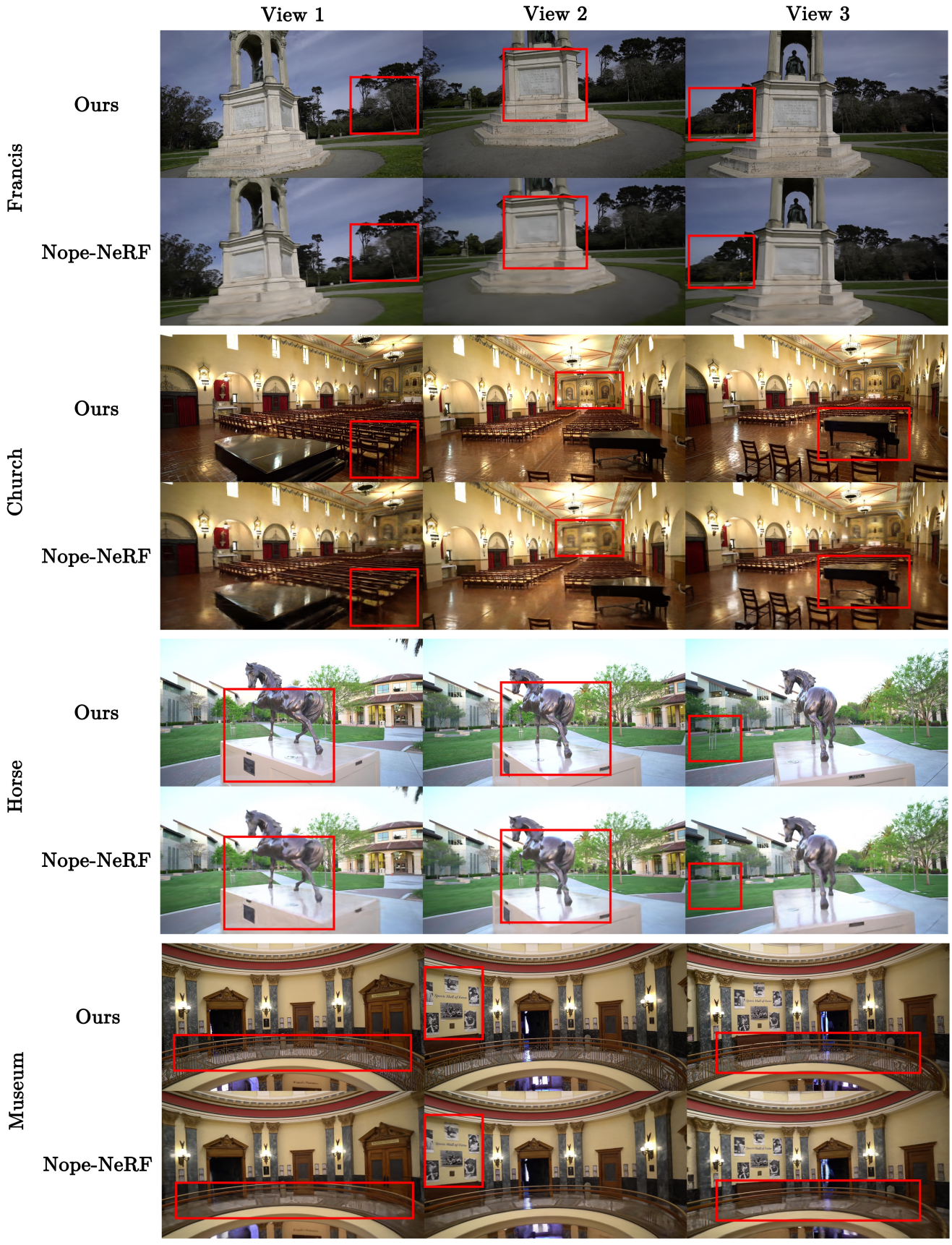

Figure 2. **Qualitative comparison for novel view synthesis on Tanks and Temples.** For each method, we fit the learned trajectory with a bezier curve and uniformly sample new viewpoints for rendering. **Better viewed when zoomed in.**

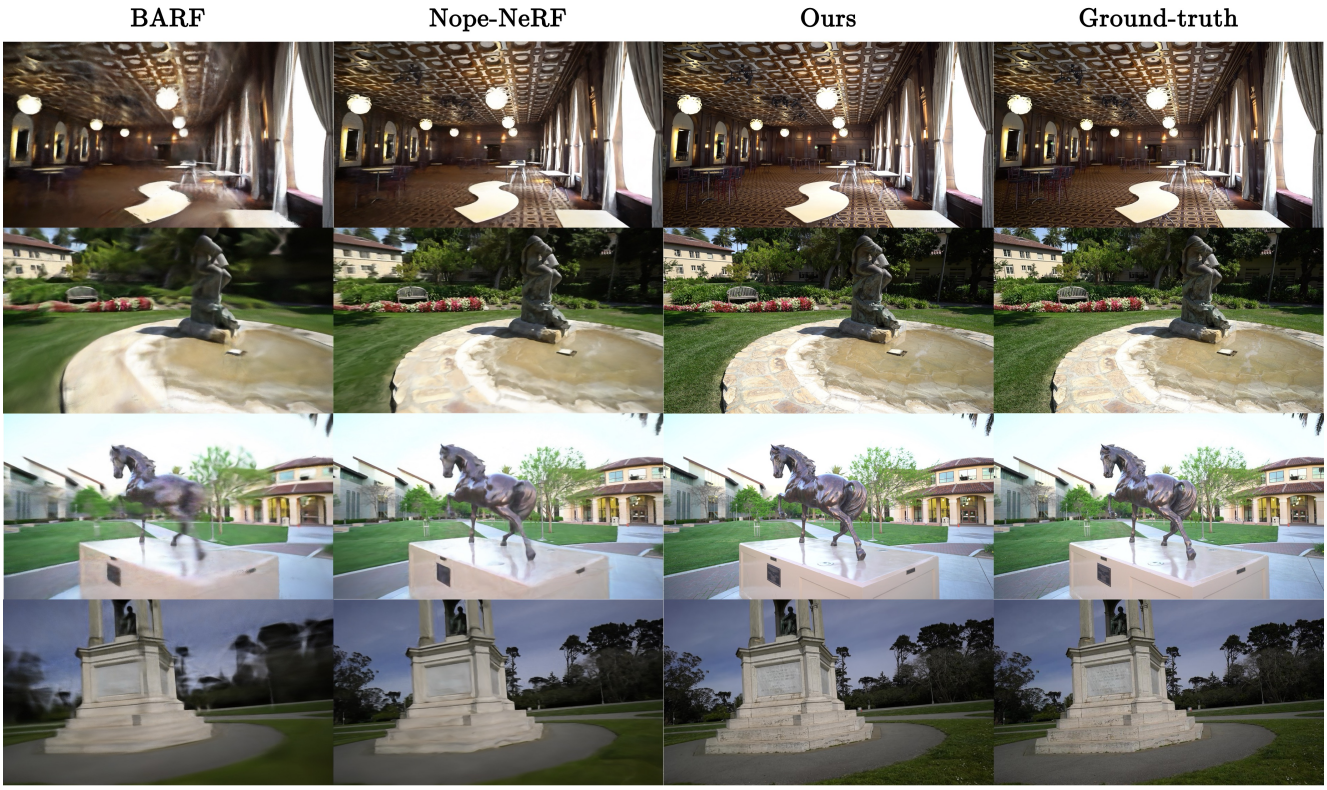

Figure 3. **Qualitative comparison for novel view synthesis on Tanks and Temples.** Our approach produces more realistic rendering results than other baselines. **Better viewed when zoomed in.**

Additional results on CO3D-V2. We conduct experiments on 5 additional scenes of the CO3D-V2 dataset and the novel view synthesis results are summarized in Table 2.

Additional Visualization. We present additional qualitative results for novel view synthesis on Tanks and Temples and CO3D-V2 in Fig. 3 and Fig. 4 following the same evaluation procedure described in the main paper.

B.2. Camera Pose Estimation

Additional results on CO3D-V2. We conduct experiments on 5 additional scenes of the CO3D-V2 dataset for the task of camera pose estimation. The results are reported in Table 3. We show better performances than Nope-NeRF [2] in both pose accuracy and synthesis quality.

Additional Visualization. Additional qualitative results for camera pose estimation on CO3D-V2 are presented in Fig. 5, following the evaluation procedure outlined in the main paper. In scenarios involving large camera motions, our approach significantly outperforms Nope-NeRF.

scenes	Nope-NeRF			Ours		
	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE
189_20393_38136	0.444	2.84	0.034	0.064	0.225	0.007
247_26441_50907	0.34	1.395	0.032	0.395	0.477	0.007
407_54965_106262	0.553	4.685	0.057	0.31	0.243	0.008
429_60388_117059	0.398	2.914	0.055	0.134	0.542	0.018
46_2587_7531	0.426	4.226	0.023	0.095	0.447	0.009
mean	0.432	3.212	0.040	**0.200**	**0.387**	**0.010**

Table 3. **Camera Pose Estimation on CO3D V2.** The best results are highlighted in bold.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1

[2] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 1, 2, 5

[3] 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] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1

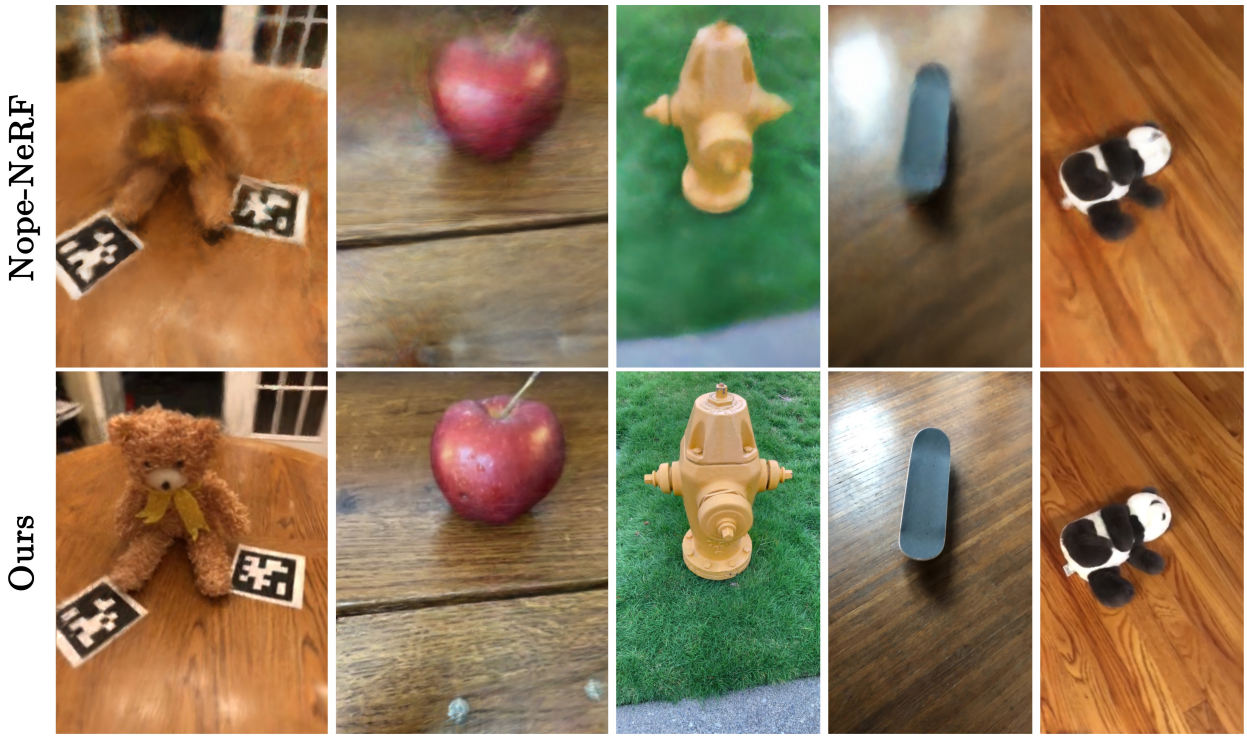

Figure 4. **Qualitative comparison for novel view synthesis on CO3D-V2.**Our approach produces more realistic rendering results than other baselines. Better viewed when zoomed in.

[5] 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. 1

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[7] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 2

[8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2

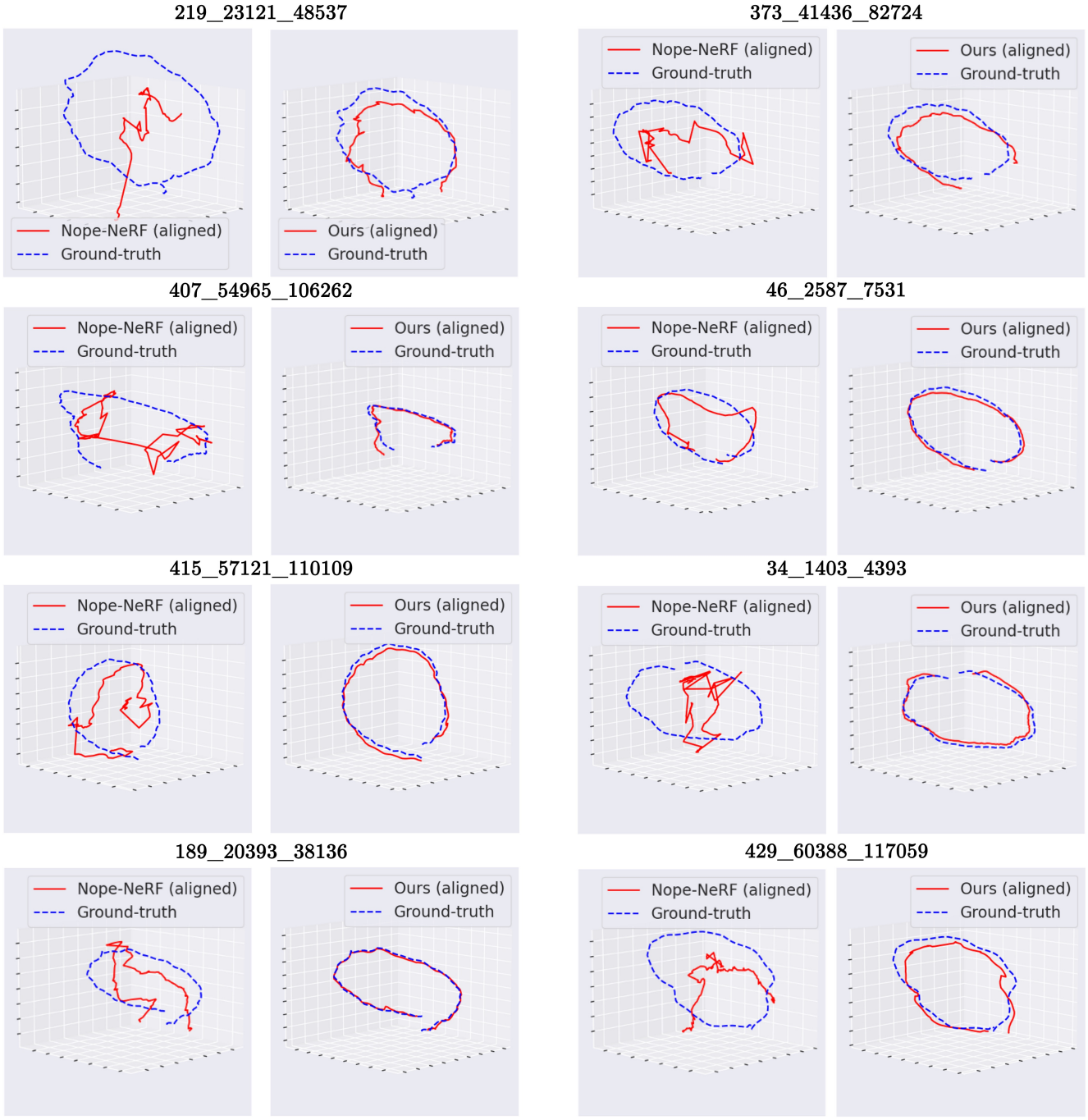

Figure 5. **Qualitative comparison for Camera Pose Estimation on CO3D-V2.** The ground-truth trajectory and the estimated one are shown in blue and red, respectively.