PCR-GS: Colmap-Free 3D Gaussian Splatting via Pose Co-Regularizations Supplementary Material

Yu Wei[1] Jiahui Zhang[1] Xiaoqin Zhang[2] Ling Shao[3] Shijian Lu[1]

[1]Nanyang Technological University [2]Zhejiang University of Technology
[3]UCAS-Terminus AI Lab, University of Chinese Academy of Sciences

1. Overview

This supplementary material provides more details and experimental results of the proposed PCR-GS. In the following sections, we present more details about the Tanks&Temples dataset, more qualitative results, ethical considerations, limitation analysis, and hyperparameter configurations.

2. Dataset

We reconstruct our Tanks&Temples dataset to create scenes with more complex trajectories from the original video as shown in Table. 1 compared with the original setting shown in Table. 2. Specifically, we sample frames from longer trajectories at much lower frame rates resulting in drastic rotation and translation of camera poses between adjacent frames. This operation introduces challenges when reconstructing 3D scenes without camera pose priors.

Besides, on the dataset split strategy, we follow the strategy used in CF-3DGS. We assess novel view synthesis quality and pose estimation accuracy on both indoor and outdoor scenes. For each scene, we use seven images per eight-frame clip for training and evaluate synthesis quality on the remaining images.

Scenes	Type	Seq. length	Fps
Church	Indoor	20s	7
Barn	Outdoor	20s	5
Museum	Indoor	30s	4
Family	Outdoor	20s	6
Horse	Outdoor	20s	6
Ballroom	Indoor	40s	1.5
Francis	Outdoor	50s	2
Ignatius	Outdoor	30s	4

Table 1. **Details of the sampling strategy** on our Tanks&Temples dataset. **Fps** refers to the number of frames per second, while **Seq. length** represents the total duration (in seconds) of the sampled video. Compared to the Tanks&Temples dataset used in CF-3DGS[1], we sampled fewer frames per second from longer camera trajectories to increase rotation and translation between adjacent frames. The average frame rate for all the scenes is 4 fps.

3. More Qualitative and Quantitative Results

We also train and test our method on CO3D-V2 [3] dataset. 2 scenes are selected from the dataset. CO3D-V2 comprises of object-centric videos, in which large and complicated camera motions make reconstruction a very challenging task. The quantitative comparision results are shown in Table. 3 and 4.

We present additional qualitative results of novel view synthesis on Tanks&Temples, Free-Dataset [4] and CO3D-V2 compared with pose-free methods including Nope-NeRF [1] and CF-3DGS [2]. As Fig. 1, Fig. 2 and Fig. 3 show, our proposed method can render images with finer structures and textures. This visualization proves that the proposed PCR-GS can effectively reconstruct 3D scenes with complex camera trajectories without relying on pose priors.

4. Ethical Consideration

While the proposed PCR-GS excels at learning photo-realistic scene representations, its capabilities could potentially be misused for illegal purposes, such as facilitating image forgery. To address such concerns, a possible solution is to incorporate watermarks into the rendered images, clearly indicating that they are synthetic.

5. Limitation

Despite the superior performance of PCR-GS, our method relies on DINO feature reprojection to regularize relative pose, which introduces inaccuracies during reprojection.

Figure 1. **More qualitative results of novel view synthesis.** We compare the results from baseline models and PCR-GS on the scenes 'Ballroom', 'Family', and 'Ignatius' from Tanks&Temples. The highlighted regions demonstrate the superior reconstruction performance of our method.

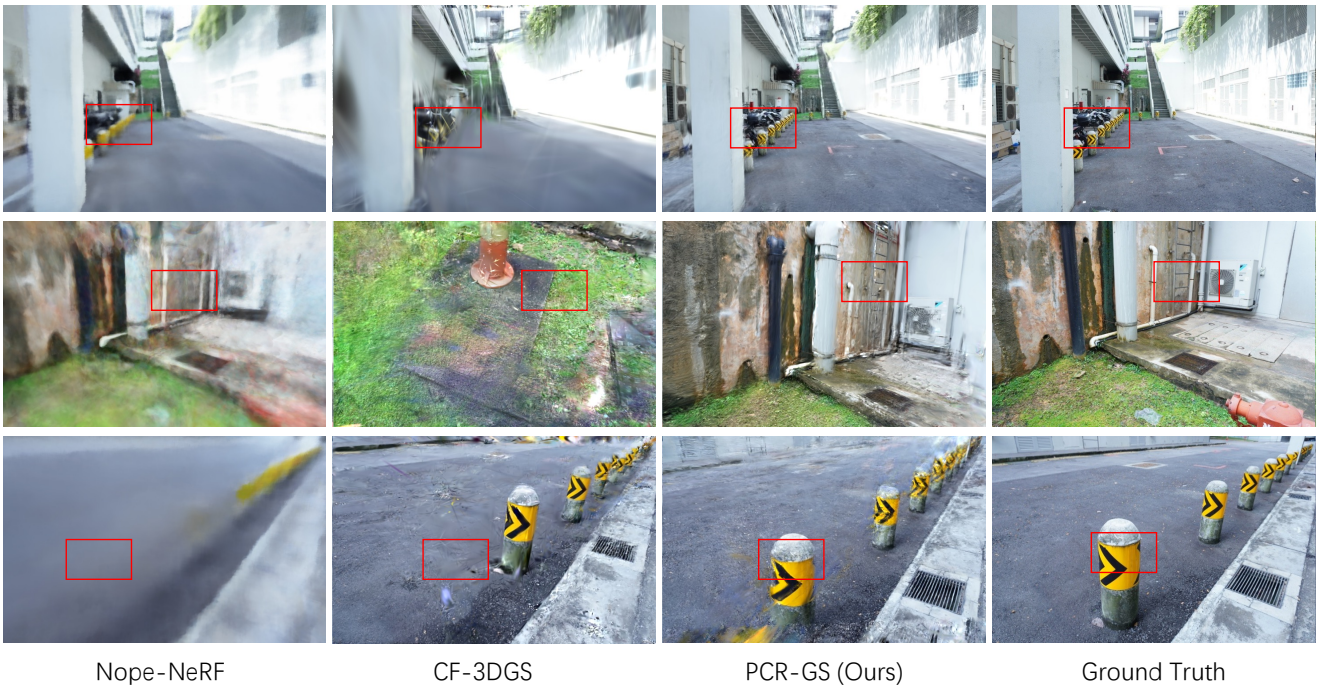

Figure 2. **More qualitative results of novel view synthesis.** We compare the results from baseline models and PCR-GS on the scenes 'stair', 'hydrant', and 'pillar' from Free-Dataset. The highlighted regions demonstrate the superior reconstruction performance of our method.

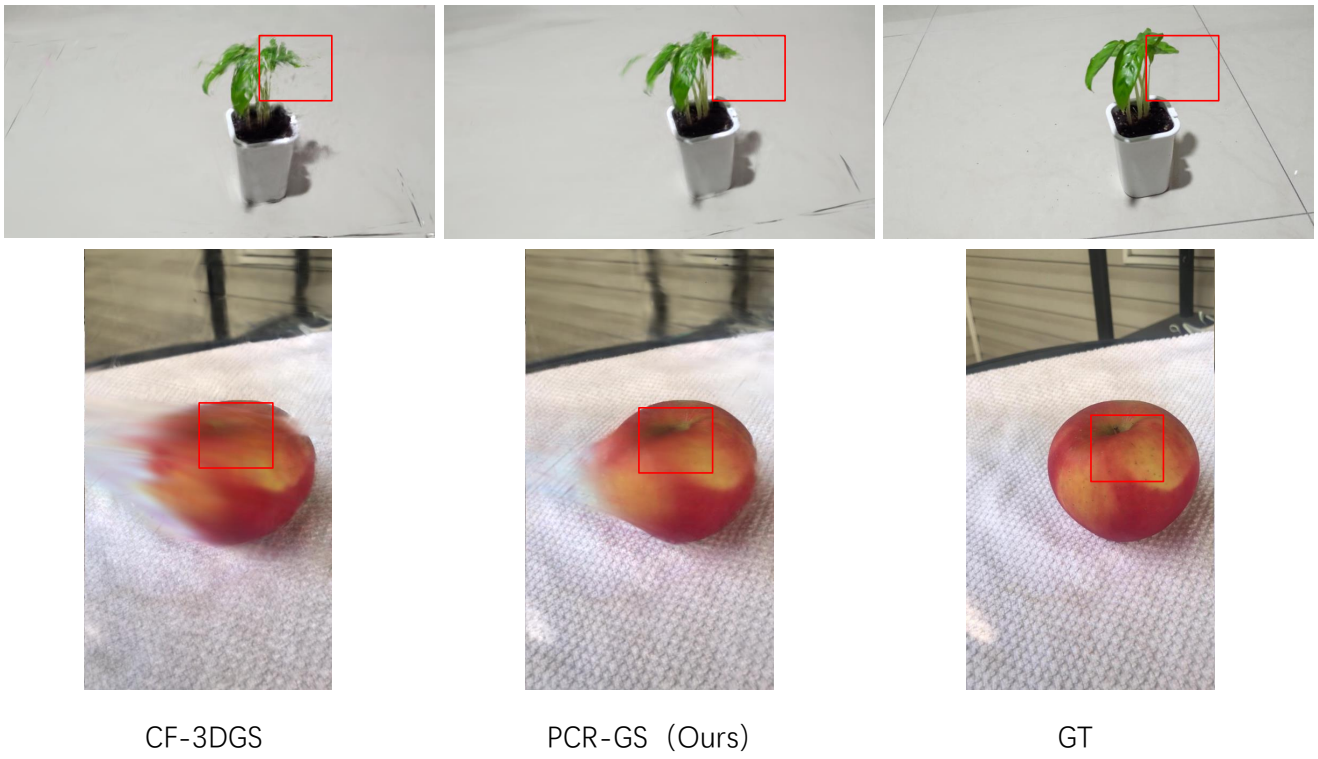

Figure 3. **More qualitative results of novel view synthesis.** We compare the results from baseline models and PCR-GS on the scenes 'Apple' and 'Plant' from CO3D-V2. The highlighted regions demonstrate the superior reconstruction performance of our method.

Scenes	Type	Seq. length	Fps
Church	Indoor	14s	29
Barn	Outdoor	15s	10
Museum	Indoor	10s	10
Family	Outdoor	6s	33
Horse	Outdoor	6s	20
Ballroom	Indoor	8s	19
Francis	Outdoor	15s	20
Ignatius	Outdoor	6s	20

Table 2. **Details of the sampling strategy** on the Tanks&Temples dataset reported in CF-3DGS. **Fps** refers to the number of frames per second, while **Seq. length** represents the total duration (in seconds) of the sampled video. The average frame rate for sampling original dataset used in CF-3DGS is 20 fps.

CO3D-V2	CF-3DGS			PCR-GS (ours)		
	RPE_t↓	RPE_r↓	ATE↓	RPE_t↓	RPE_r↓	ATE↓
Apple	**0.541**	0.571	0.020	0.544	**0.569**	**0.019**
Plant	0.626	1.765	0.038	**0.538**	**1.241**	**0.021**
Mean	0.584	1.168	0.029	**0.541**	**0.905**	**0.020**

Table 3. Quantitative comparisons on pose estimation over CO3D-V2. The best score is in bold.

CO3D-V2	CF-3DGS			PCR-GS (ours)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Apple	**18.53**	0.68	0.43	18.50	**0.72**	**0.40**
Plant	20.46	0.91	0.29	**20.59**	**0.94**	**0.26**
Mean	19.50	0.80	0.36	**19.55**	**0.83**	**0.33**

Table 4. Quantitative comparisons on novel view synthesis over CO3D-V2. The best score is in bold.

As the foundation of reprojection, the depth value of each pixel is rendered from 3D Gaussians pre-trained on a single frame, which may lead to scale and shift ambiguity. Such depth bias may limit the accuracy of the feature reprojection. Moving forward, we will investigate a better way to provide precise and unbiased depth prediction before reprojection operation.

6. Hyperparameter Configurations

We leverage DINO feature reprojection regularization and wavelet-based frequency regularization to regularize the camera pose during the training process. The weight of each regularization term in the loss function is determined through extensive experimentation. Following the configuration in CF-3DGS [2], the RGB regularization is composed of a L_1 loss and a D-SSIM of the rendered image and the ground truth:

$$\mathcal{L}_{\text{rgb}} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}}. \tag{1}$$

We use $\lambda = 0.2$ for all experiments. Meanwhile, we set the weights of the feature reprojection regularization term and wavelet-based frequency regularization term in the total loss function to 0.2. The total loss function is illustrated in Eq. 2.

$$\mathcal{L}_{\text{total}} = \lambda_0\mathcal{L}_{\text{RGB}} + \lambda_1\mathcal{L}_{\text{Feat}} + \lambda_2\mathcal{L}_{\text{Freq}}, \tag{2}$$

where $\mathcal{L}_{\text{RGB}}$ is the RGB regularization term as defined in Eq. 1, while $\mathcal{L}_{\text{Feat}}$ and $\mathcal{L}_{\text{Freq}}$ denote the feature regularization term and frequency regularization term, respectively. The weights for these terms are set as $\lambda_0 = 0.6$, $\lambda_1 = \lambda_2 = 0.2$.

References

[1] 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] 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*, pages 20796–20805, 2024. 1, 4

[3] 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

[4] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4150–4159, 2023. 1