ManifoldNeuS: Manifold-aware View Optimizability for Pose-Free Neural Surface Reconstruction (Supplementary Material)

Xinxin Liu Xue Wang Guoqing Zhou Qing Wang *
School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
liuxxin26@mail.nwpu.edu.cn, {xwang, zhouguoqing, qwang}@nwpu.edu.cn

1. Content

In this supplementary material, we present extended experiments and implementation details to complement the main paper. Specifically, the supplementary material includes the following contents:

(i) Additional implementation details covering the training details, hyperparameter settings, and model architecture (Sec. 2).

(ii) Analysis of the easy-view bias in the uniform view optimization (Sec. 3).

(iii) More experiments, including results on additional scenes from the DUT dataset, evaluations on novel view synthesis, comparisons against pose-free surface reconstruction methods, performance under sparse-view settings, and computational efficiency comparisons (Sec. 4).

(iv) Further ablation studies analyzing hyperparameter sensitivity and key design choices of our method, including the balancing factor in MaVOS, comparisons with alternative view scheduling strategies and different coverage metrics for MaVOS, the number of views used in the view scheduling, components of MaVOS-gated positional encoding, and different loss function combinations (Sec. 5).

These supplementary results further clarify the proposed method and provide additional evidence of its effectiveness and robustness in pose estimation and scene reconstruction.

2. Implementation Details

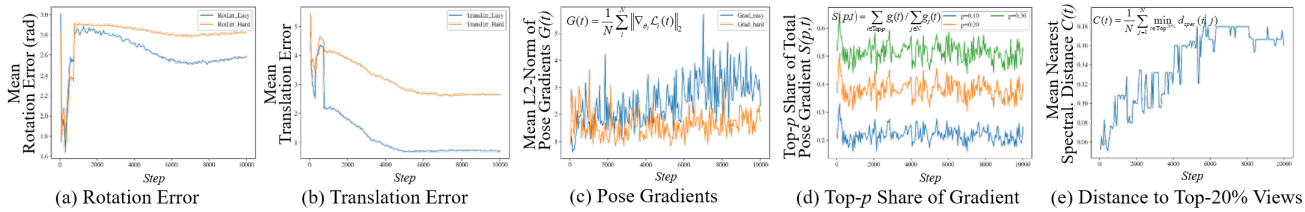
We employ two separate optimizers for scene representation and pose optimization, respectively. For scene modeling, we adopt the Adam optimizer [3] with a learning rate that warms up linearly from 0 to 5×10^{-4} over the first 5k iterations, then decays to 2.5×10^{-5} following a cosine annealing schedule. For pose estimation, we use the AdamW optimizer [6] with base learning rates of 8×10^{-3} and 5×10^{-3} for the anchors' pose estimation stage and the remaining estimation stage, respectively. Its learning rate also undergoes a linear warm-up from zero across the initial 2k iterations, after which it decays via cosine annealing, gradually reducing to $0.1\times$ the base learning rate by the end of training. For the hyperparameter settings, we select the top 30% of views with the highest cumulative feature matching count as anchor views, and choose 20 successor views in each remaining optimization round. For the positional encoding, the maximum number of frequency bases for sampling position and view direction is set to 6 and 4, respectively. The offset scaling factor in Eq. 9 of the main manuscript is set to 0.3. For the model architecture, we utilize two MLPs to approximate geometry and appearance, detailed in Eq. 1 of the main manuscript, following NeuS [10]. We use SIFT descriptors [7] for feature correspondences.

3. Analysis of Easy-View Bias

To our knowledge, we are the first to identify easy-view bias in pose-free neural surface reconstruction. We support it through an analysis of NeuS-BARF with a uniform view optimization by comparing the performances of easy views and hard views as shown in Fig. A1.

It can be seen that, easy views, defined as the top 20% ranked by view overlap, consistently exhibit lower rotation and translation errors than hard views (bottom 20%) throughout training, as shown in Fig. A1 (a-b), revealing inherent optimization bias. Moreover, easy views produce substantially larger pose-gradient norms in Fig. A1 (c), suggesting that they dominate the optimization updates under uniform training process. This effect is further quantified in Fig. A1 (d), where the top-ranked easy views account for a large fraction of the total pose-gradient norm (e.g., the top 20% easy views contribute over 40%), providing evidence of local gradient dominance. Meanwhile, Fig. A1 (e) shows that the mean nearest spectral distance from all views to these top-ranked views decreases slowly during training and even increases, meaning global coverage ne-

*Corresponding author.



(a) Rotation Error (b) Translation Error (c) Pose Gradients (d) Top-p Share of Gradient (e) Distance to Top-20% Views

Figure A1. Analysis of easy-view bias in uniform training using NeuS-BARF. We compare the differences between easy views (top 20% ranked by view overlap) and hard views (bottom 20%) in terms of: (a) mean rotation error, (b) mean translation error, (c) mean L_2 norm of pose gradients, (d) Top-p share of the total pose-gradient norm, and (e) normalized mean nearest spectral distance from all views to the top-20% easy views during training processing of *Scan24*.

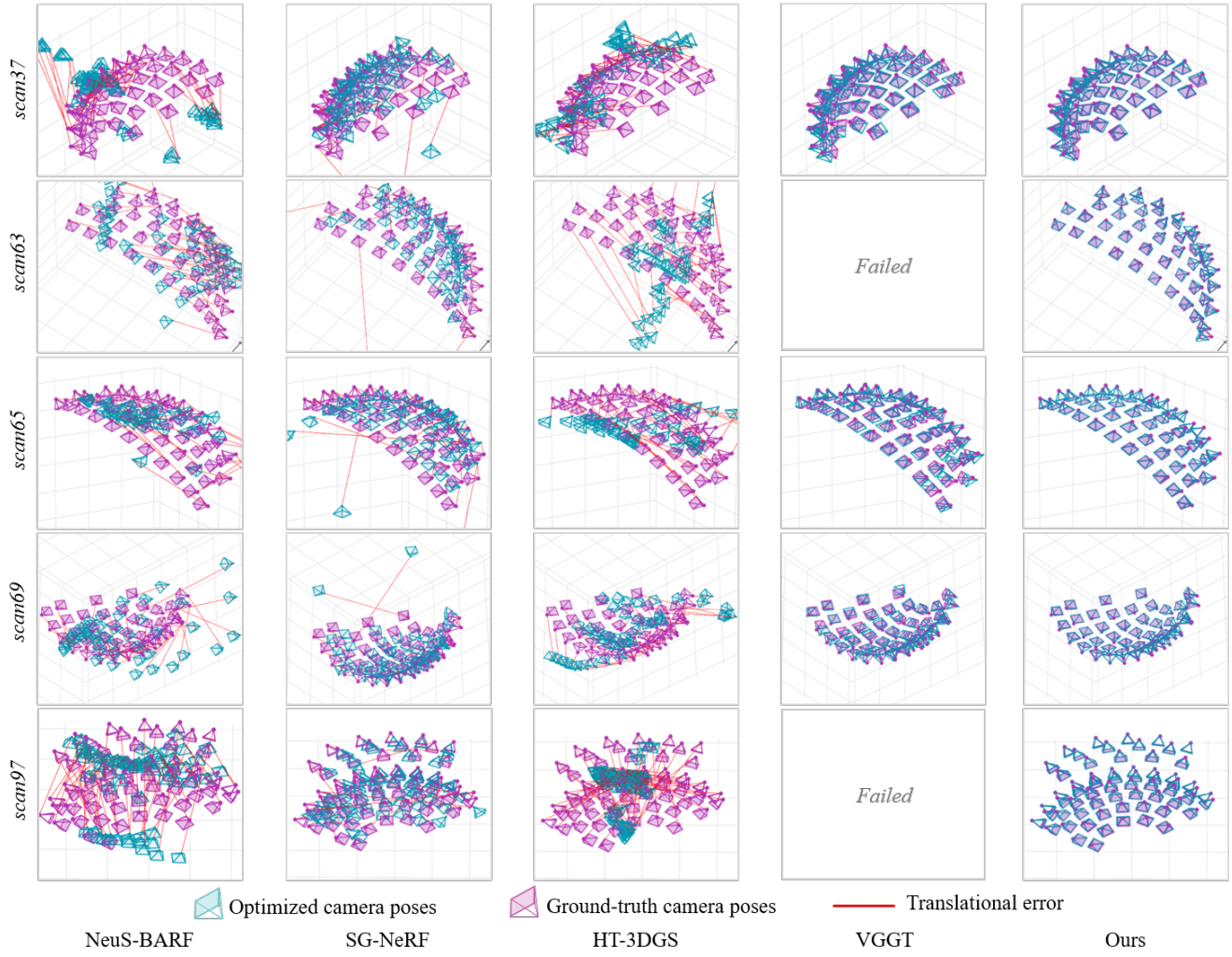


Figure A2. Visual comparison of the ground-truth and optimized camera poses (Procrustes aligned). Our method successfully realigns the camera frames, while baselines converge to suboptimal or noisy estimates. The results of VGGT are post-processed by bundle adjustment.

glect. Together, these results reveal a fundamental imbalance in uniform training: optimization is biased toward easy views that dominate gradient updates, and views critical for global topological coverage are underemphasized. To address this issue, we introduce MaVOS in Eq. 6 (main paper) to formalize view optimizability through two complementary components: 1) topological coverage (worth optimizing) measured by spectral-embedding distance to mitigate

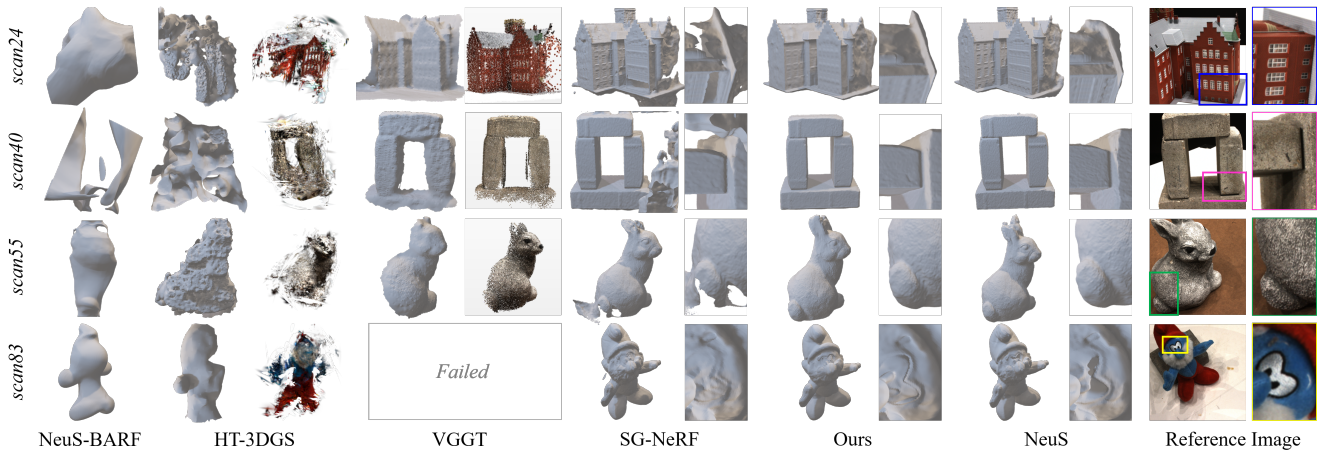


Figure A3. Qualitative comparisons of reconstruction quality. For each method, the geometric mesh extracted via Marching Cubes [5] is shown on the left. On the right side of each column, we present the denoised results of the original representations for HT-3DGS and VGGT. Specifically, the results of VGGT are postprocessed by bundle adjustment. Additionally, zoomed-in views for SG-NeRF, NeuS (trained with COLMAP poses), and our method are provided to highlight fine geometric details. Zoom in for the best view.

bias, and 2) local constraint (easy to optimize) measured by feature-level co-visibility to guide convergence.

4. More Experiments and Results

In this section, we present additional experimental results on the DUT dataset, along with further evaluations in terms of novel view synthesis, sparse-view performance, and computational efficiency, as well as comparisons with pose-free surface reconstruction methods.

4.1. More Results on DTU

We provide additional results for the remaining scenes not included in the main manuscript. Qualitative comparisons of pose estimation and reconstruction quality are shown in Fig. A2 and Fig. A3, respectively.

NeuS-BARF and HT-3DGS fail to reconstruct reasonable geometries across all scenes, primarily due to highly inaccurate pose estimates, as evidenced in Fig. A2. VGGT cannot produce a valid reconstruction for *Scan83*; similar to *Scan97*, this scene contains multi-scale structures, as illustrated by the camera distributions in Fig. 2 of the main manuscript and in Fig. A2, underscoring VGGT's difficulty in handling scenes with significant scale variation. SG-NeRF broadly recovers overall object shapes but exhibits blurred boundaries and missing components—particularly in *Scan24* and *Scan55*—and introduces noticeable artifacts in *Scan40* and *Scan83*, as visualized in Fig. A3. This degradation stems from its estimation of noisy or inconsistent poses. NeuS trained with COLMAP-provided poses achieves accurate geometry thanks to the reliability of the camera poses. Our method consistently outperforms all baselines in terms of detail preservation and surface coherence. As confirmed by the zoomed-in views, our method effectively preserves geometric details while suppressing noise and fragmentation. These results demonstrate that our method not only surpasses existing pose-free methods but also rivals pose-supervised baselines, validating its effectiveness in achieving high-fidelity neural surface reconstruction without relying on known camera poses.

4.2. Novel View Synthesis

Accurate geometry recovery enables high-fidelity novel view synthesis. To further validate the performance of our method in pose-free reconstruction, we compare our method against SG-NeRF [1], HT-3DGS [2], VGGT [9], and NeuS [10] on novel view synthesis using the recovered geometry with predicted camera poses. For evaluation, we hold out 1 in 8 of all images per scene as the test set.

From the visual quality comparison illustrated in Fig. A4, it can be observed that SG-NeRF produces rendered views with noticeable artifacts (*e.g.*, in *Scan24*) due to its noisy reconstructed geometry, while inaccurate pose estimates further lead to blurring, as seen on the wall in *Scan24* and the brick surface in *Scan37*. HT-3DGS, due to reconstruction failures, exhibits significant structural distortion and texture loss in its rendered images, along with characteristic Gaussian artifacts. Since VGGT is designed purely for reconstruction rather than novel view synthesis, we apply bundle adjustment to its raw noisy point cloud output following [9] and perform Poisson surface reconstruction to obtain a geometry mesh. Given a camera pose of the test view, we then project the mesh into a 2D image to obtain the rendered view. However, the post-processed point clouds become too sparse, resulting in blurry renderings. Moreover, as shown in *Scan24* and *Scan65*, the discrete

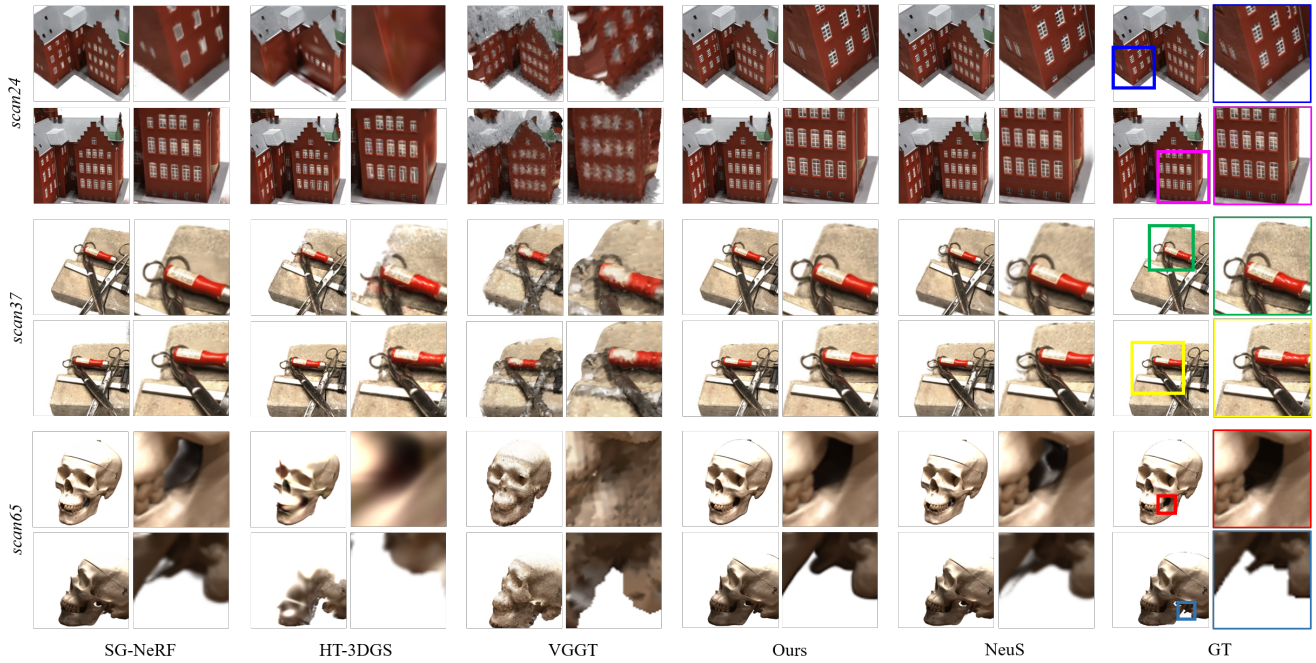


Figure A4. Qualitative results of novel view synthesis. The ground-truth images are processed using the dataset-provided masks. For each scene, we provide results from two test views, with zoomed-in regions shown on the right for detailed evaluation.

Table A1. Quantitative comparisons of novel view synthesis. PSNR, SSIM, and LPIPS scores are reported for the evaluation of novel views. The best scores are shown in bold, and the second-best results are underlined. Our method achieves comparable performance to NeuS.

Scan	SG-NeRF [1]			HT-3DGS [2]			VGGT [9]			Ours			NeuS [10]		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
24	25.52	0.805	0.218	23.17	0.782	0.280	14.72	0.540	0.505	<u>29.07</u>	<u>0.905</u>	<u>0.102</u>	**29.64**	**0.918**	**0.098**
37	<u>28.89</u>	<u>0.901</u>	<u>0.107</u>	23.91	0.887	0.119	16.52	0.642	0.309	**29.03**	**0.906**	**0.100**	27.82	0.880	0.128
65	<u>28.45</u>	<u>0.968</u>	<u>0.095</u>	19.71	0.848	0.301	15.99	0.733	0.279	**28.83**	**0.974**	**0.084**	28.22	0.967	0.095
Avg.	27.62	0.891	0.140	22.26	0.839	0.233	15.75	0.638	0.364	**28.98**	**0.928**	**0.095**	<u>28.56</u>	<u>0.922</u>	<u>0.107</u>

nature of the point cloud leads to missing geometry in the mesh, and noisy point clouds introduce structural distortion in the synthesized views, as shown in *Scan37*. In contrast, our method faithfully reconstructs both scene geometry and high-frequency surface details, producing novel views with visual fidelity. The quantitative comparison results presented in Table A1 further demonstrate our method's superiority in preserving fine geometric details and rendering photorealistic novel views.

4.3. Comparisons with Pose-Free Surface Reconstruction Methods

We further compare our method with representative pose-free surface reconstruction methods NoPose-NeuS [8] and ParaSurRe [4] on DTU. Since the official implementations of them are not publicly available, we follow their reported evaluation settings and report the corresponding results from their papers for comparison.

We compare against ParaSurRe under its reported DTU evaluation setting, which includes four scans, three of which are beyond our original evaluation subset, *e.g.*, *Scan106*, *Scan114*, and *Scan122*. Under this protocol, our method achieves lower mean rotation and translation errors than ParaSurRe, i.e., 0.34/0.77 versus 0.37/1.11. For NoPose-NeuS, we conduct a matched comparison on their DTU scenes using the same metrics, namely relative rotation error, relative translation error, and Chamfer distance (PRE_r/PRE_t/CD). The results are summarized in Table A2. Compared with NoPose-NeuS, our method achieves better overall performance, further validating its effectiveness for pose-free surface reconstruction.

Table A2. Comparisons with pose-free surface reconstruction method NoPose-NeuS [8]. We report relative rotation and translation errors, as well as Chamfer distance. The best scores are shown in bold. Zoom in for the best view.

	24	37	40	55	63	65	69	83	97	106	114	122	Avg.
NoPose	0.55/1.01/0.91	0.89/1.08/1.51	0.69/0.91/0.95	0.54/0.88/0.44	0.66/**0.76**/1.01	0.54/**0.80**/0.63	0.62/0.87/0.79	0.79/1.02/1.53	0.57/1.19/1.22	0.48/0.79/0.51	0.60/0.96/**0.39**	0.58/0.97/0.67	0.63/0.94/0.88
Our	**0.32/0.87/0.34**	**0.67/0.83/0.37**	**0.46/0.81/0.42**	**0.22/0.72/0.34**	**0.54**/0.79/**0.55**	**0.30**/0.84/**0.47**	**0.30/0.56/0.36**	**0.51/0.87/0.40**	**0.29/0.64/0.39**	**0.34/0.64/0.43**	**0.31/0.77**/0.54	**0.43/0.81/0.62**	**0.39/0.76/0.44**

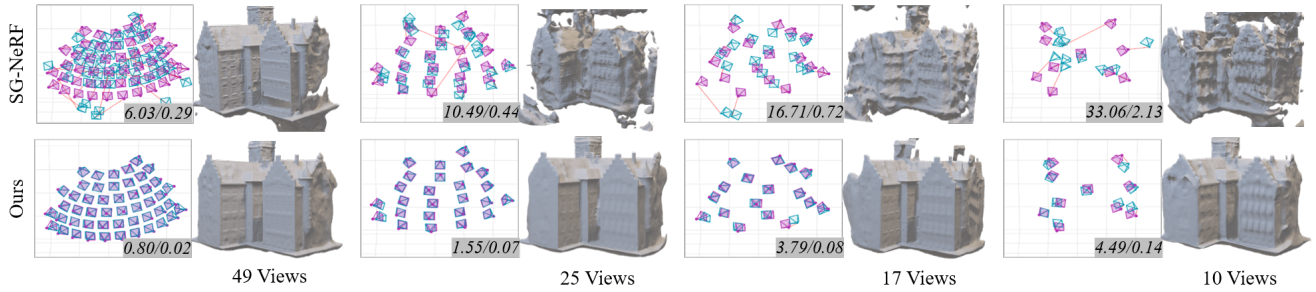


Figure A5. Comparisons under low co-visibility and sparse input in *Scan24*. We visualize pose estimates and geometry, and report quantitative pose errors ($\Delta R/\Delta T$) highlighted in gray.

4.4. Comparisons under Sparse-View Settings

We evaluate the robustness of our method to challenging view distributions. We reduce view overlap by subsampling the original 49 views at intervals of 2, 3, and 5, yielding 25, 17, and 10 views. Fig. A5 shows our method outperforms SG-NeRF in pose accuracy and geometric fidelity, even with only 10 views, demonstrating robustness to low co-visibility and sparse input.

4.5. Computational efficiency

Our average training time is 6.5h, faster than SG-NeRF (8h) and comparable to NeuS-BARF (6h), while achieving better pose and geometry. It is slightly slower than NeuS (5.2h) since we additionally optimize camera poses. HT-3DGS (0.5h) and VGGT (3min) are faster but yield inferior pose/reconstruction quality (as shown in the main text). Overall, our method offers a favorable trade-off of accuracy and efficiency.

5. Additional Ablation Studies

In this section, we provide further ablation experiments on the following hyperparameters and design choices, including: (i) the balancing hyperparameter in MaVOS, (ii) comparisons with alternative view scheduling strategies and different coverage metrics for MaVOS, (iii) the number of anchor and successor views in the view scheduling, (iv) the three terms in the MaVOS-gated PE module, and (v) different combinations of loss terms. Our experimental results primarily focus on pose estimation, as camera poses fundamentally determine the quality of the final 3D reconstruction. We perform the expanded ablation studies on the anchor and the first successor views optimization rounds, denoted as "Anc. Opt." and "Succ. Opt." respectively.

Table A3. Ablation study of the hyperparameter α of MaVOS on metrics rotation ($\triangle R$) and translation ($\triangle T$ errors). We report the comparative results of anchor and successor optimization rounds at various hyperparameter values. When α takes larger values (*e.g.*, 1.00 or 0.85), MaVOS predominantly relies on feature co-visibility based local constraints strength, whereas smaller values prioritize topological coverage. Bold indicates the results of the default setting in our method.

α	$\triangle R \downarrow$		$\triangle T \downarrow$	
	Anc. Opt.	Succ. Opt.	Anc. Opt.	Succ. Opt.
1.00	5.735	0.348	1.563	0.086
0.85	4.305	0.251	1.530	0.070
0.65	**3.883**	**0.225**	**1.076**	**0.050**
0.50	6.500	0.401	1.191	0.061
0.35	9.800	0.577	2.281	0.105

This focused evaluation strategy is justified because subsequent successor optimization rounds employ identical implementation parameters and algorithms as the first successor round, making additional ablations redundant for evaluating key method components.

Effect of balancing hyperparameter in MaVOS. The hyperparameter α in Eq. 6 of the main manuscript controls the trade-off between local constraint strength and global topological coverage. We conduct experiments with various values of α.

Table A3 quantifies the influence of parameter α on pose optimization performance across anchor and successor phases. Experimental results reveal a distinct performance trade-off: higher α values prioritize local constraints, yielding lower pose errors during anchor optimization but compromising performance in successor stages. This behavior

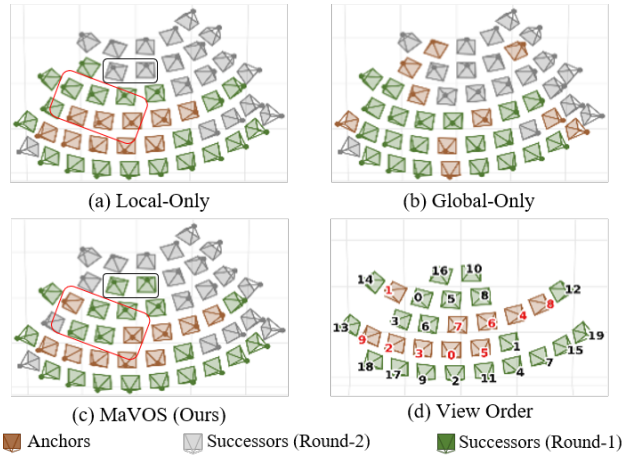


Figure A6. Visualization of view scheduling. We compare the anchor and successor selections produced by MaVOS using only local constraint strength (a), only global topological coverage (b), and the full formulation (c). (d) shows the view ranking (high-to-low score) obtained by MaVOS. Brown, green, and gray indicate anchors, first-round successors, and second-round successors, respectively.

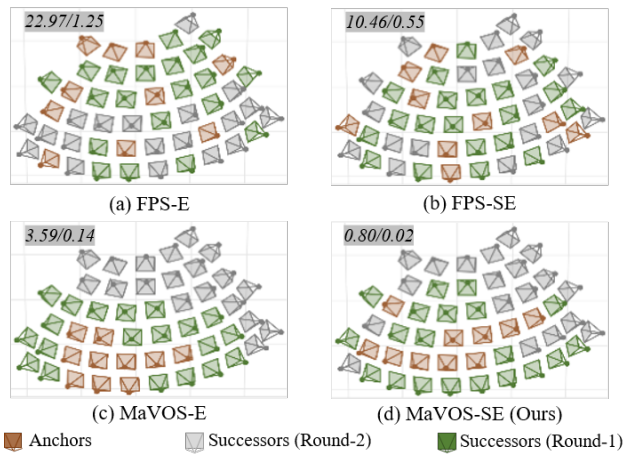


Figure A7. Comparisons of alternative view scheduling strategies and coverage metrics. We compare FPS and MaVOS using Euclidean distance on the co-visibility matrix (E) and spectral-embedding distance (SE). Pose errors ($\Delta R/\Delta T$) are reported in the upper-left corner.

stems from the tension between immediate geometric fitting quality and long-term coverage optimization. The former requires strong local constraints for stable pose initialization from identity transformations, while the latter demands sufficient view diversity for global consistency. Conversely, lower α values prioritize topological coverage at the cost of anchor pose accuracy, creating suboptimal initialization for successors. The configuration $\alpha = 0.65$ achieves an optimal balance, maintaining competitive anchor accuracy while enabling effective successor optimization through appropriate topological guidance. This balanced setting delivers the most consistent performance across both stages, making it our default parameter.

We also visualize MaVOS's ability to balance local constraint strength and global topological coverage by observing its influence on view scheduling. Fig. A6 visualizes the anchor and successor views selected by different view scheduling settings, together with the view ranking. Using only local constraints, *i.e.*, $\alpha = 1.0$, as shown in Fig. A6 (a), tends to favor views with strong overlap, resulting in anchors spatially concentrated in a clustered region (red box) and leaving large unexplored regions (black box) for successors (round-2). Although such a strategy benefits early optimization stability, it lacks global exploration. Using only global coverage, *i.e.*, $\alpha = 0.0$, as shown in Fig. A6 (b), produces a much more scattered distribution of selected views. This improves spatial spread, but many selected anchors are globally diverse and weakly overlapping with one another, which cannot provide insufficient support for reliable pose estimations in the early stages. In contrast, MaVOS in Fig. A6 (c) produces a balanced selection with better global connectivity and sufficient local support.

Comparisons with alternative view scheduling and different coverage metrics. We compare our MaVOS with farthest point sampling (FPS), and further ablate the coverage metric used in MaVOS. Both FPS and MaVOS are evaluated with two distance choices: Euclidean distance on the co-visibility matrix (E) and our distance in the spectral embedding space (SE). From the quantitative comparisons reported in Fig. A7, FPS-E and FPS-SE produce larger pose errors than the corresponding MaVOS variants, respectively. This is because FPS only maximizes view dispersion and ignores local constraint strength, often yielding weakly connected views with insufficient overlap for stable optimization, as shown in Fig. A7 (a-b). Comparing MaVOS-E and MaVOS-SE in Fig. A7 (c–d), MaVOS-SE (ours) achieves topology-aware coverage, while topology-agnostic MaVOS-E shows over-clustered view selection, indicating spectral embedding's necessity for global connectivity. Overall, these results show that both local constraint modeling and topology-aware global coverage are necessary for effective view scheduling.

Effect of the number of views in view scheduling. We investigate the influence of the number of anchor views used during the anchor optimization. For fixed-anchor pose predictions, we further analyze how varying the number of views affects performance in subsequent successor optimization.

As summarized in Table A4, the parameter K represents the percentage of views selected as anchors based on cumulative feature correspondences, where views are prioritized according to their local constraint strength. As expected,

Table A4. Ablation study of anchor view percentage ($K\%$) and successor views per optimization round (M). Analysis of K focuses exclusively on the anchor optimization, while M evaluation is confined to the successor optimization under consistent anchor pose estimates. Bold indicates the results of the default settings in our method.

	Varying Top $K\%$ Anchor Views					Varying M Successor Views				
	$K=10$	$K=20$	$K=30$	$K=40$	$K=50$	$M=10$	$M=15$	$M=20$	$M=25$	$M=30$
$\triangle R\downarrow$	2.941	2.811	**3.883**	5.117	6.517	0.938	1.073	**1.076**	1.386	1.620
$\triangle T\downarrow$	0.166	0.162	**0.225**	0.306	0.373	0.027	0.044	**0.050**	0.052	0.062

increasing view counts generally increases pose errors due to optimization complexity. Nevertheless, during anchor optimization, increasing K from 20% to 30% only yields a marginal increase in error, suggesting that the top 30% of views provide an effective balance of good initialization without significant accuracy degradation. Similarly, in the successor optimization, using $M=20$ views achieves pose accuracy comparable to that of $M=15$. Since reducing the number of views per optimization round increases the total number of iterations required for convergence, we select $K=30\%$ and $M=20$ as the optimal configuration, balancing performance and efficiency.

Effect of the key terms in MaVOS-guided PE. We conduct ablation experiments on the three key terms, *i.e.*, $\widetilde{S}_i$, η_f, and η_t in Eq. 8 of the main manuscript to assess their individual effectiveness. Furthermore, we evaluate the impact of different values of the scaling factor λ. Results are detailed in the Table A5.

From experimental results, a key finding is that the base model yields lower errors in the anchor stage, but it brings no significant improvement in the subsequent successor stage. This observation indicates that the anchor set primarily consists of views with abundant textual features, resulting in marginal differences in view optimizability among them. In contrast, the successor optimization incorporates a larger set of views (including both initial anchors and current successors), which amplifies the optimizability disparity and exacerbates the issue of easy-view-bias. Our MaVOS-guided PE successfully mitigates successor-stage errors, providing compelling evidence that explicitly modeling and incorporating per-view optimizability within PE is essential.

In the ablation study of positional encoding components, frequency modulation is more effective at reducing both rotation and translation errors. This highlights the critical role of frequency-aware modulation in preserving high-frequency geometric details. Meanwhile, temporal modulation contributes to performance by modestly enhancing optimization stability. Furthermore, during the successor optimization, these two modulation terms enable more accurate translation estimation. For the scaling factor, while $\lambda=0.1$ yields optimal anchor phase performance, $\lambda=0.3$ achieves superior successor optimization, providing the optimal balance between initial stability and progressive optimization capability.

Table A5. Ablation study on positional encoding components. "Base" represents the original BARF PE without our MaVOS-guided strategy. $\widetilde{S}_i$ denotes our PE with only per-view optimizability guidance. We introduce two key sensitivity parameters: η_f, which captures frequency sensitivity to enforce stronger modulation on higher-frequency components, and η_t, which controls temporal sensitivity to gradually intensify the modulation effect over the optimization progress. Performance is measured by pose accuracy in both the anchor (Anc. Opt.) and successor (Succ. Opt.) optimization rounds. Best and second-best results are highlighted in bold and underlined separately for ablation studies of components and scaling factor λ.

Setting	$\triangle R\downarrow$		$\triangle T\downarrow$	
	Anc. Opt.	Suc. Opt.	Anc. Opt.	Suc. Opt.
Modulations				
Base	**3.206**	3.574	**0.158**	0.152
+ $\widetilde{S}_i$	5.150	1.981	0.282	0.084
+ $\widetilde{S}_i$ + η_f	4.156	<u>1.351</u>	0.248	<u>0.067</u>
+ $\widetilde{S}_i$ + η_t	4.560	1.662	0.271	0.076
Offset scale				
$\lambda=0.1$	**3.858**	1.432	<u>0.219</u>	<u>0.066</u>
$\lambda=0.3$	4.402	<u>1.416</u>	0.261	<u>0.066</u>
$\lambda=1.0$	4.874	1.934	0.275	0.831
Full model				
$0.3\cdot\widetilde{S}_i\cdot\eta_f\cdot\eta_t$	<u>3.883</u>	**1.076**	<u>0.225</u>	**0.050**

Table A6. Ablation study of the losses. "Recon." denotes the losses used in NeuS reconstruction, including photometric loss, Eikonal loss, and mask loss. "Depth" and "Normal" denote the addition of depth loss and normal loss into "Recon." losses, respectively.

Losses	$\triangle R\downarrow$		$\triangle T\downarrow$	
	Anc. Opt.	Succ. Opt.	Anc. Opt.	Succ. Opt.
Recon.	9.741	7.545	0.600	0.467
Depth	7.271	6.661	0.452	0.413
Normal	4.113	2.073	0.289	0.076
All	**3.883**	**1.076**	**0.225**	**0.050**

Effect of losses. Table A6 presents an ablation study of the loss functions used in our method. The normal loss substantially contributes to geometric accuracy across both anchor ("Anc. Opt.") and successor ("Succ. Opt.") optimization rounds, yielding measurable reductions in rotational error ($\triangle R$) and positional drift ($\triangle T$). The depth loss primarily reduces rotational error. Overall, these results demonstrate that normal constraints provide foundational geometric supervision essential for the optimizations, whereas depth supervision serves as a complementary signal primarily benefiting rotational precision.

References

[1] Yiyang Chen, Siyan Dong, Xulong Wang, Lulu Cai, Youyi Zheng, and Yanchao Yang. SG-NeRF: Neural surface reconstruction with scene graph optimization. In *European Conference on Computer Vision (ECCV)*, pages 188–205. Springer, 2024. 3, 4

[2] Bo Ji and Angela Yao. SfM-free 3D gaussian splatting via hierarchical training. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 21654–21663, 2025. 3, 4

[3] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[4] Wenyu Li, Zongxin Ye, Sidun Liu, Ziteng Zhang, Xi Wang, Peng Qiao, and Yong Dou. ParaSurRe: Parallel surface reconstruction with no pose prior. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 4

[5] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer graphics*, pages 163–169, 1987. 3

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[7] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1

[8] Mohamed Shawky Sabae, Hoda Anis Baraka, and Mayada Mansour Hadhoud. NoPose-NeuS: Jointly optimizing camera poses with neural implicit surfaces for multi-view reconstruction. *arXiv preprint arXiv:2312.15238*, 2023. 4, 5

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 5294–5306, 2025. 3, 4

[10] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, pages 27171–27183, 2021. 1, 3, 4