High-Resolution and Few-shot View Synthesis from Asymmetric Dual-lens System ——Supplementary Material——

Ruikang Xu[1], Mingde Yao[2], Yue Li[1], Yueyi Zhang[1], Zhiwei Xiong[1]✉

[1] MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
xurk@mail.ustc.edu.cn, zwxiong@ustc.edu.cn
[2] The Chinese University of Hong Kong

The supplementary material is organized as follows:

- Section 1 provides ablations for multi-reference-guided refinement.
- Section 2 provides visual results of ablations.
- Section 3 provides results of comparison with baselines using telephoto.
- Section 4 provides comparison results without SR task.
- Section 5 provides more implementation details.
- Section 6 provides comparisons of rendered depth.
- Section 7 provides additional visual comparisons.

1 More Ablations for Multi-reference-guided Refinement

1.1 Ablation on loss function terms

After 3D-GS optimization, we can directly synthesize wide-angle views at the target resolution through pre-upsampling, but they still lack realistic high-frequency details due to the resolution limitation. We design a multi-reference-guided refinement module to refine the newly synthesized view, generating the reconstructed result $\hat{I}^{rec}$. The proposed module is trained by a self-training loss, since ground-truth data is inaccessible in practical development, denoted as

$$\mathcal{L}_{DL} = \lambda_1 \|\text{Crop}(\hat{I}^{rec}) - T^{align}\|_2 + \lambda_2 \|\hat{I}^{rec} - I^{\uparrow}\|_2 + \lambda_3 \mathcal{L}_{cx}(\text{Crop}(\hat{I}^{rec}), T). \quad (1)$$

We denote the first term as $\mathcal{L}_{tele}$, which leverages HR information from the spatially aligned image T^{align}. T^{align} is pre-aligned from $T \in \{T_i\}_{i=1}^{M}$ using optical flow estimation. We denote the second term as $\mathcal{L}_{wide}$, which provides an essential content constraint using the pre-upsampled image $I^{\uparrow} \in \{I_i^{\uparrow}\}_{i=1}^{M}$. The third term is the contextual loss $\mathcal{L}_{cx}$ [5,6], which measures the similarity without considering the spatial positions to maximize HR information transfer. Crop$(\cdot)$ denotes the cropping operation to obtain the overlapped FoV area of two lenses.

To analyze the impact of the different terms of the self-training loss function, we perform an ablation study on the simulated dataset as shown in Table 1. The results highlight the significance of $\mathcal{L}_{tele}$ and $\mathcal{L}_{wide}$ in improving the reconstruction fidelity, while $\mathcal{L}_{cx}$ contributes to perceptual metric improvement.

Table 1: Ablation on the terms of the proposed self-training loss (Eq. 1), using the simulated dataset with the 90-shot training samples while a scale factor is 2×.

$\mathcal{L}_{tele}$	$\mathcal{L}_{wide}$	$\mathcal{L}_{cx}$	PSNR ↑	SSIM ↑	LPIPS ↓
✓	✗	✗	25.42	0.7655	0.2101
✓	✓	✗	25.58	0.7679	0.2094
✓	✓	✓	25.61	0.7692	0.2076

1.2 Ablation on selected dual-lens pair numbers

To refine the high-frequency details of newly synthesized views, we propose the multi-reference-guided refinement module, which uses the formulation of exploiting the information of guided images to reconstruction [4, 8]. However, the corresponding telephoto images to the newly synthesized views are unavailable in practical applications, which results in that we can not directly obtain the corresponding guided images. To address this problem, we propose to select guided images from the dual-lens training samples based on the camera pose distances. Specifically, we select the telephone images to exploit their HR information while selecting the wide-angle images to utilize their multi-view information.

To analyze the influence of the selected dual-lens pair numbers on performance, we conduct an ablation study as shown in Table 2. We can see that the performance improves with an increasing number of selected pairs, but it diminishes after reaching 2 pairs. This observation aligns with practical considerations, where more dual-lens pairs provide more information for reconstruction but the incremental benefit becomes limited. Therefore, the number of dual-lens pairs is set to 2, achieving a balance between computational resources and performance.

Table 2: Ablation on the number of selected dual-lens image pairs, using the simulated dataset with the 90-shot training samples while a scale factor is 2×.

Number	1-pair	2-pair	3-pair	3-pair	5-pair
PSNR ↑	25.43	25.61	25.70	25.72	25.75
SSIM ↑	0.7647	0.7692	0.7704	0.7708	0.7710
Param (M)	0.72	0.81	0.96	1.09	1.25

2 Visual Results of Ablations.

We supplement the visual results for the ablation of two key components, *i.e.*, consistency-aware training strategy (CTS) and multi-reference-guided refinement (MR), and the ablation of the losses used in MR. As shown in Fig. 1, the top half demonstrates the effectiveness of our proposed components in improving reconstruction fidelity. The bottom half shows that $\mathcal{L}_{wide}$ provides the content constraint to imperfect telephoto image alignment and $\mathcal{L}_{cx}$ facilitates HR information transfer.

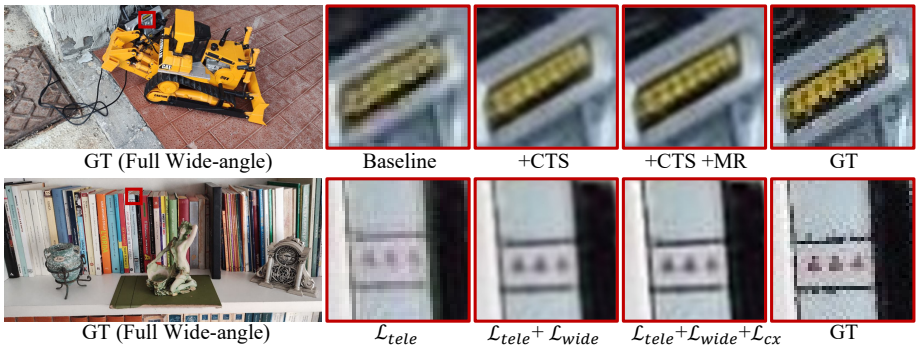

Fig. 1: Visual results of ablation studies.

3 Comparison with Baselines using Telephoto Images

3.1 Baselines directly taking telephoto as input

To further verify the effectiveness of our method, we retrain 3D-GS [3] and SparseNeRF [7] with both wide-angle and telephoto as inputs (termed with †) in Table 3 and Fig. 2. It can be seen that the results drop heavily since NeRF/GS-based methods cannot directly handle the training samples collected with different intrinsic parameters, which breaks the assumption of camera uniqueness [2]. That is the motivation of our method design.

Table 3: Comparison on the forward-facing scenes of the real-captured dataset with the 5-shot samples and 2× SR.

Method	3D-GS+DCSR	3D-GS†+DCSR	SparseNeRF+DCSR	SparseNeRF†+DCSR	Ours
PSNR ↑	21.03	15.29	23.19	15.95	24.05
SSIM ↑	0.7099	0.5046	0.7142	0.6185	0.7525

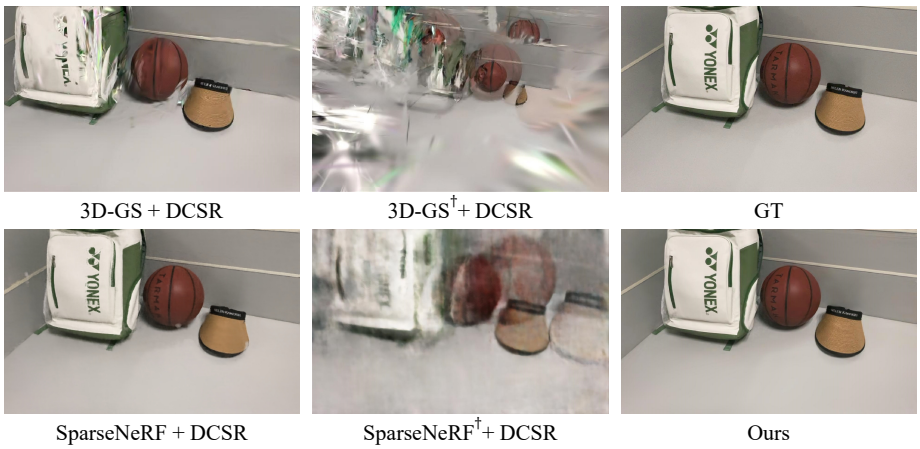

Fig. 2: Visual comparison with NVS methods taking telephoto as input.

3.2 Baselines using telephoto images by a two-stage strategy

To enable NVS baselines to utilize the high-resolution information from telephoto images, we integrate these methods with the dual-lens SR method (*i.e.*, DCSR [8]). Beyond 3D-GS [3] + DCSR [8] compared in the main manuscript, we further compare our method with SparseNeRF [7] and FSGS [9] followed by DCSR [8] in Table 4, demonstrating the general superiority of our method.

Table 4: Comparison on the forward-facing scenes of the real-captured dataset with the 5-shot samples and 2× SR.

Method	SparseNeRF+HAT	FSGS+HAT	SparseNeRF+DCSR	FSGS+DCSR	Ours
PSNR ↑	22.98	23.08	23.19	23.32	24.05
SSIM ↑	0.7127	0.7322	0.7142	0.7333	0.7525

4 Comparison without SR

To further verify the effectiveness of our proposed two regularization terms ($\mathcal{R}_c$ and $\mathcal{R}_d$), we reproduce our method without the pre-upsampling and the SR part and compare it with SparseNeRF [7] and FSGS [9]. Results in Table 5 demonstrate the superiority of our method, even when SR is not introduced.

Table 5: Comparison without SR on the forward-facing scenes of the real-captured dataset with the 5-shot samples.

Method	3D-GS [3]	SparseNeRF [7]	FSGS [9]	Ours w/o SR
PSNR ↑	22.90	25.11	25.26	26.52
SSIM ↑	0.7383	0.7451	0.7529	0.7694

5 More Implementation Details

Setup of COLMAP. For the real-world dataset, we run COLMAP over the full dataset to estimate accurate camera parameters, and then sample a few of them as input. For the simulated dataset, following DS-NerRF [1], we directly run COLMAP on the training images (10/20/90 shots).

Resolutions of input/output/train/test images. Resolutions vary across different scenes. In simulated and real-world datasets, output resolutions are around 1160×522. The output resolution is 2× the input resolution. The training and testing sets have the same image resolutions.

6 Comparison of Rendered Depth

In this section, we provide visual comparisons of rendered depth maps from different solutions. Specifically, we perform the comparison on the few-shot training cases on the simulated and real-captured datasets in Fig. 3 and Fig. 4.

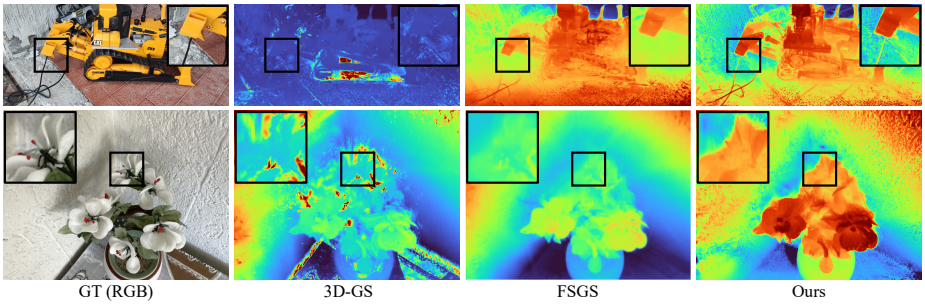

Fig. 3: Visual comparison of rendered depth maps with 3D-GS-based solutions for the few-shot training case (10-shot) on the simulated dataset.

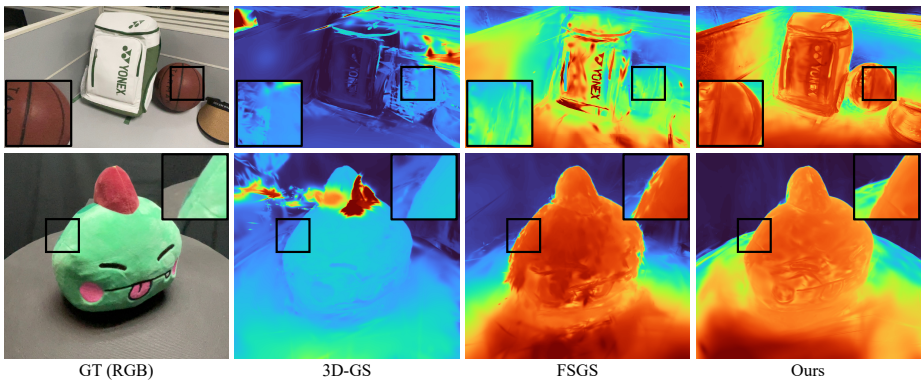

Fig. 4: Visual comparison of rendered depth maps with 3D-GS-based solutions for the few-shot training cases (5-shot for the forward scene on the top half, 15-shot for the inward scene on the bottom half) on the real-captured dataset.

7 More Results of Visual Comparison

More visual results on the real-captured dataset can be found below.

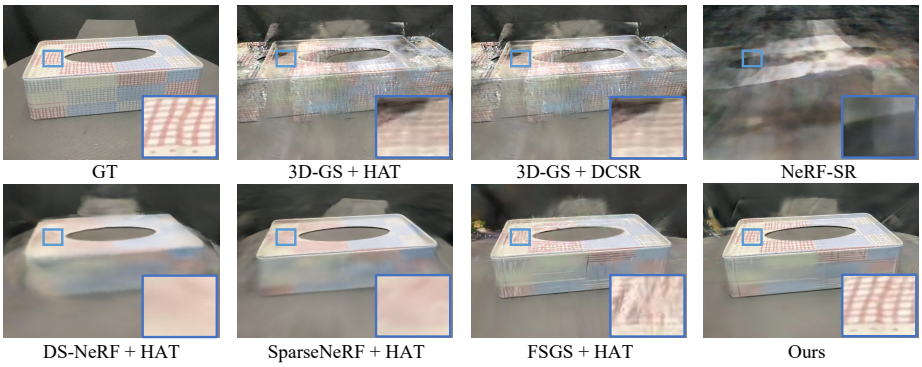

Fig. 5: Visual comparison with different solutions using the 15-shot training sample on the inward scene of the real-captured dataset.

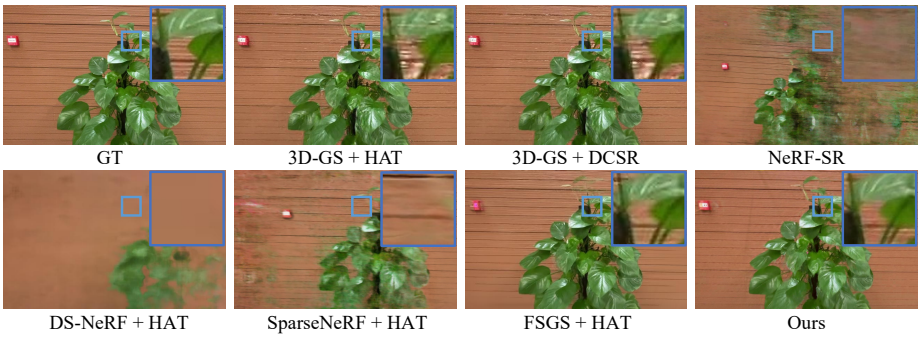

Fig. 6: Visual comparison with different solutions using the 5-shot training sample on the forward scene of the real-captured dataset.

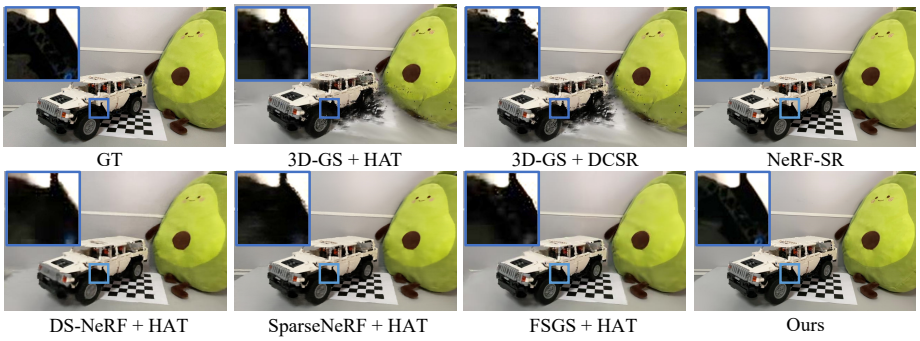

Fig. 7: Visual comparison with different solutions using the 50-shot training sample on the forward scene of the real-captured dataset.

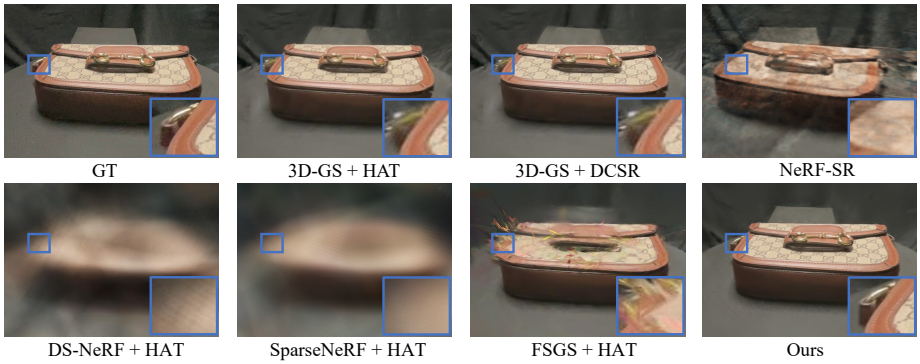

Fig. 8: Visual comparison with different solutions using the 50-shot training sample on the inward scene of the real-captured dataset.

References

1. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: CVPR (2022) 4

2. Gao, Y., Su, L., Liang, H., Yue, Y., Yang, Y., Fu, M.: Mc-nerf: Multi-camera neural radiance fields for multi-camera image acquisition systems. arXiv preprint arXiv:2309.07846 (2024) 3
3. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023) 3, 4
4. Lee, J., Lee, M., Cho, S., Lee, S.: Reference-based video super-resolution using multi-camera video triplets. In: CVPR (2022) 2
5. Mechrez, R., Talmi, I., Shama, F., Zelnik-Manor, L.: Maintaining natural image statistics with the contextual loss. In: ACCV (2019) 1
6. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: ECCV (2018) 1
7. Wang, G., Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In: ICCV (2023) 3, 4
8. Wang, T., Xie, J., Sun, W., Yan, Q., Chen, Q.: Dual-camera super-resolution with aligned attention modules. In: ICCV (2021) 2, 4
9. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. arXiv preprint arXiv:2312.00451 (2023) 4