

AERGS-SLAM: Auto-Exposure-Robust Stereo 3D Gaussian Splatting SLAM

Supplementary Material

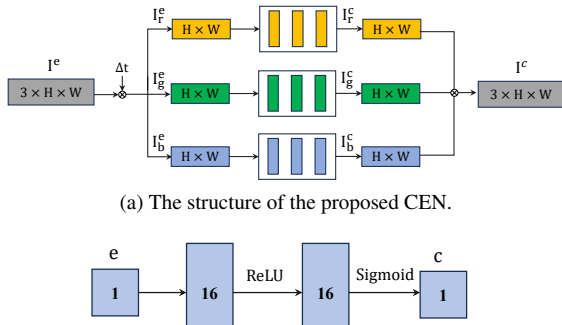
A. Overview

We provide extensive qualitative results in **the attached local webpage (accessed by Demo.html)**. The supplementary material is organized as follows: Section B provides details of Camera Exposure Network (CEN); Section C reports the real-time performance; Section D provides details of the Evaluation Dataset; Section E provides additional ablation results; Section F provides additional experimental results.

B. Details of CEN

The proposed CEN maps per-image radiance maps to RGB images. We use three independent Multi-Layer Perceptrons (MLPs) to model this mapping process.

Firstly, as shown in Fig. 1a, the radiance map \mathbf{I}^e is split into three channels, forming three independent radiance maps \mathbf{I}_r^e , \mathbf{I}_g^e , and \mathbf{I}_b^e . Meanwhile, three independent MLPs are employed to model the mapping from the radiance map of each channel to color maps. This mapping corresponds to the camera response function (CRF). As shown in Fig. 1b, we use a lightweight MLP to model the CRF. The input is the radiance value of each pixel, and each MLP consists of two hidden layers with an embedding size of 16, using ReLU activation. The output is the color value, with Sigmoid activation applied.



(b) The structure of each MLP for mapping radiance e to color c .

Figure 1. illustration of the CEN.

Secondly, the exposure-rendering equation is

$$\mathbf{I}^c = g(\ln \sum_{i \in N} \mathbf{e}_i \cdot G'_i \alpha_i \prod_{j=1}^{i-1} (1 - G'_j \alpha_j) + \ln \Delta t), \quad (1)$$

where Δt is exposure time. Notably, the logarithmic exposure time $\ln \Delta t$ here leads to scale ambiguity of the ex-

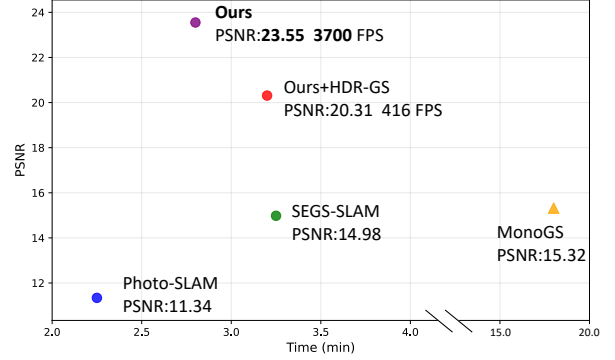


Figure 2. The real-time performance of the proposed method and baselines

posure time, i.e., $\ln \alpha \Delta t = \ln \Delta t + \ln \alpha$, where $\ln \alpha$ can be modeled by 3D Gaussians. Therefore, we only recover the relative exposure time. For the evaluation of camera exposure estimation, we utilize the real exposure times to fit a scale factor and a bias to recover the absolute exposure times.

C. Real-time performance

As shown in Fig. 2, we report the real-time performance metrics of the proposed method and baselines, including runtime (minutes), rendering quality (PSNR, higher is better), and exposure-controlled rendering speed (FPS). Firstly, the proposed method achieves the highest rendering quality with shorter runtime, benefiting from the lightweight CEN that efficiently maps radiance maps to RGB images. In contrast, HDR-GS [2] performs per-Gaussian mapping during optimization, which significantly increases the system's runtime. Similarly, SEGS-SLAM [7] integrates appearance embeddings to predict all Gaussian parameters in the optimization process, which degrades real-time performance. More importantly, AE-induced image appearance variations cannot be overcome by such view-dependent appearance embeddings.

Secondly, compared with HDR-GS, the proposed method achieves extremely fast exposure-controlled rendering. Specifically, it runs nearly 10× faster than HDR-GS. It benefits from the fact that it only needs to render the radiance map once to enable arbitrary exposure control based on the CRF.

Finally, compared with Photo-SLAM [4], the proposed method achieves a significant improvement in photorealistic mapping quality at the cost of only a slight sacrifice in real-time performance. Furthermore, the decoupled ar-

chitecture adopted in our work significantly outperforms MonoGS with a coupled framework significantly in terms of real-time performance.

D. Details of the Evaluation Dataset

We provide details of the stereo EuRoC dataset [1] and our self-collected datasets. The EuRoC MAV dataset [1] used for evaluation is processed by adjusting image brightness to simulate AE-induced exposure variations. Our self-collected dataset is captured using a Zed2i stereo camera across six challenging scenarios (i.e., S1 to S6), containing 1,947, 1,519, 1,391, 1,802, 2,424, and 1,191 images, respectively. Real exposure times were also recorded during data collection. Fig. 3 shows typical scenarios in the evaluation dataset.

E. Additional Ablation Results

We present per-scene ablation results of the proposed illumination-robust localization (IRL), coarse-to-fine optimization (CTFO), and camera exposure network (CEN) modules. As shown in Table 1, by adding the IRL module (i.e., the row (2) of Table 1), the localization and photorealistic mapping performance is significantly improved. The PSNR, SSIM, LPIPS, and RMSE metric improve notably, confirming its illumination robustness. Then, adding the CTFO module (i.e., the row (3) of Table 1) further refines photorealistic mapping performance. Integrating the CEN module into the full method (i.e., the row (5) of Table 1) achieves combined enhancement, achieving the top result in most sequences across all metrics and outperforming all ablation baselines in most scenes. Thus, the proposed method is effective under exposure-varying scenarios.

F. Additional Experimental Results

In this section, we provide more quantitative and qualitative results of illumination-robust Localization (i.e., Section F.1), exposure estimation (i.e., Section F.2), and novel view synthesis (i.e., Section F.3).

F.1. Illumination-robust Localization

Firstly, per-scene RMSE metrics of the EuRoC MAV stereo dataset [1] are reported in Table 2. We observe that the proposed method comprehensively outperforms other 3DGS-based methods on almost all sequences, as they rely on handcrafted features for localization. This further validates the advantage of the learning-based features adopted in our work.

Secondly, for qualitative validation, per-scene trajectory results are presented in Figs. 4 and 5. We also observe that the proposed method can accurately estimate the camera’s trajectory, whereas other handcrafted-feature-based methods exhibit significant trajectory drift under complex cam-

era motions. Specifically, in the trajectory comparison results of various methods for scenes V202, V103, V203, S2, S3, and S6 in Figs. 4 and 5, the proposed method (i.e., the red line) is more consistent with the ground truths (i.e., the black line).

Finally, we conclude that the proposed method achieves superior performance in both RMSE (localization accuracy) and trajectory estimation, which fully demonstrates the effectiveness of the adopted learning-based features in improving localization robustness and trajectory accuracy under exposure-varying scenarios.

F.2. Exposure Estimation

Firstly, per-scene estimated CRF curves of the EuRoC dataset and our self-collected dataset are shown in Figs. 6 and 7. First, as shown in Fig. 6, the CRF curves of all scenes are well estimated. Moreover, the CRF curves of all scenes exhibit almost identical shapes, which is physically consistent with the uniqueness of the camera’s CRF. Second, Fig. 7 comprehensively demonstrates the CRF and exposure time estimation results on our self-collected dataset. Compared with HDR-GS [2], the proposed method can estimate the camera’s CRF curve and exposure time more accurately in real-world scenes, thus further validating the effectiveness of the proposed CEN in CRF and exposure time estimation.

Secondly, more qualitative results of exposure-controlled rendering are presented in Figs. 8 and 9. The proposed method can accurately model image appearances under different exposure times: when the relative exposure time gradually increases from 0.4 s to 1.6, the image brightness increases accordingly, and our results are more consistent with real-world image characteristics compared with HDR-GS.

Finally, the proposed method demonstrates superior performance in both CRF curve and exposure time estimation. It outperforms HDR-GS [2] in exposure-controlled rendering, fully validating the effectiveness of the proposed CEN in addressing exposure-varying scenarios.

F.3. Novel View Synthesis

The per-scene photorealistic mapping results of the EuRoC dataset are presented in Table 3. Compared with other methods, the proposed method demonstrates superior comprehensive performance. Moreover, Fig. 10 provides additional qualitative results. We observe that the proposed method not only achieves high-fidelity photorealistic mapping but also faithfully adapts to scene exposure variations. In contrast, other methods (SEGS-SLAM [7], Photo-SLAM [4]) fail to model such exposure dynamics, leading to significant brightness mismatches between their rendered images and the ground truths.

Method	Metric	Sequence										
		MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203
(1) w/o CFTO, CEN, IRL	PSNR↑	12.02	12.27	11.62	12.20	13.07	12.03	11.34	X	11.86	11.06	10.17
	SSIM↑	0.341	0.356	0.386	0.441	0.467	0.545	0.577	X	0.529	0.541	0.547
	LPIPS↓	0.465	0.475	0.542	0.509	0.471	0.557	0.588	X	0.487	0.560	0.592
	RMSE↓	0.029	0.037	0.035	0.213	0.037	0.079	0.080	X	0.060	0.064	1.001
(2) w/o CTFO, CEN	PSNR↑	13.08	13.14	15.31	16.16	15.31	13.82	14.39	14.37	15.79	15.89	15.04
	SSIM↑	0.469	0.464	0.579	0.645	0.601	0.650	0.701	0.696	0.701	0.736	0.718
	LPIPS↓	0.559	0.556	0.392	0.381	0.457	0.431	0.410	0.432	0.313	0.293	0.393
	RMSE↓	0.022	0.018	0.023	0.056	0.058	0.033	0.041	0.029	0.019	0.022	0.473
(3) w/o CEN	PSNR↑	14.90	13.21	15.44	16.31	15.60	14.90	15.12	14.66	15.96	14.97	15.07
	SSIM↑	0.539	0.467	0.580	0.656	0.612	0.684	0.728	0.708	0.709	0.709	0.720
	LPIPS↓	0.444	0.532	0.363	0.358	0.419	0.361	0.347	0.403	0.267	0.323	0.369
	RMSE↓	0.022	0.018	0.023	0.057	0.071	0.033	0.020	0.031	0.047	0.022	0.218
(4) w/o CTFO	PSNR↑	16.73	16.36	19.95	20.49	18.85	22.45	23.74	21.99	22.50	22.53	19.74
	SSIM↑	0.547	0.534	0.657	0.732	0.686	0.782	0.827	0.816	0.779	0.796	0.768
	LPIPS↓	0.489	0.513	0.339	0.318	0.397	0.237	0.236	0.285	0.223	0.260	0.343
	RMSE↓	0.022	0.018	0.023	0.057	0.045	0.033	0.021	0.028	0.019	0.022	0.570
(5) Ours	PSNR↑	19.92	19.19	19.59	20.55	19.00	22.34	23.55	22.93	22.04	22.37	20.68
	SSIM↑	0.654	0.627	0.651	0.729	0.693	0.784	0.832	0.840	0.781	0.791	0.743
	LPIPS↓	0.318	0.331	0.317	0.311	0.350	0.249	0.204	0.231	0.199	0.250	0.352
	RMSE↓	0.021	0.017	0.023	0.056	0.045	0.033	0.051	0.024	0.036	0.023	0.215

Table 1. Per-scene ablation results in stereo EuRoC MaV dataset [1]. For photometric mapping metrics, we color code each column as **best**. Notably, only the IRL module contributes to the localization accuracy, i.e., the comparison of the first and the second row. 'X' denotes experimental failure.

Method	Sequences											
	MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203	Avg.
ORB-SLAM3 [3]	0.044	0.047	X	0.176	0.133	0.088	0.088	X	0.257	0.125	1.522	0.276
DROID-SLAM [6]	0.012	0.012	0.022	0.093	0.041	0.037	0.012	0.019	0.018	0.010	0.055	0.030
AirSLAM [8]	0.022	0.018	0.023	0.057	0.071	0.033	0.020	0.031	0.047	0.022	0.218	0.051
MonoGS [5]	0.089	0.065	1.821	3.505	3.134	0.115	0.421	0.745	0.280	1.592	X	1.178
Photo-SLAM [4]	0.029	0.037	0.035	0.213	0.037	0.079	0.080	X	0.060	0.064	1.001	0.164
SEGS-SLAM [7]	0.037	0.038	0.052	0.047	X	0.089	0.161	0.288	0.237	0.062	X	0.112
Ours	0.021	0.017	0.023	0.056	0.045	0.033	0.051	0.024	0.036	0.023	0.215	0.049

Table 2. Localization results (RMSE ↓) in stereo EuRoC MaV dataset [1]. We color code each column as **best** and **second best**. All others are obtained in our experiments. 'X' denotes experimental failure.



Figure 3. Visualization of selected scenarios from the evaluation dataset.

Method	Metric	Sequence											
		MH01	MH02	MH03	MH04	MH05	V101	V102	V103	V201	V202	V203	Avg.
MonoGS [5]	PSNR↑	19.07	19.62	14.99	17.15	18.47	17.43	15.32	14.90	16.93	12.86	X	16.67
	SSIM↑	0.757	0.785	0.587	0.688	0.730	0.827	0.750	0.730	0.760	0.634	X	0.725
	LPIPS↓	0.255	0.215	0.475	0.387	0.361	0.270	0.472	0.569	0.315	0.632	X	0.395
Photo-SLAM [4]	PSNR↑	12.02	12.27	11.62	12.20	13.07	12.03	11.34	X	11.86	11.06	10.17	11.76
	SSIM↑	0.341	0.356	0.386	0.441	0.467	0.545	0.577	X	0.529	0.541	0.547	0.473
	LPIPS↓	0.465	0.475	0.542	0.509	0.471	0.557	0.588	X	0.487	0.560	0.592	0.525
SEGS-SLAM [7]	PSNR↑	15.99	16.96	16.34	18.21	X	15.23	14.98	15.51	15.45	14.14	X	15.87
	SSIM↑	0.621	0.653	0.654	0.741	X	0.712	0.730	0.748	0.712	0.716	X	0.699
	LPIPS↓	0.324	0.279	0.274	0.224	X	0.267	0.327	0.332	0.253	0.272	X	0.284
Ours + HDR-GS [2]	PSNR↑	18.99	19.42	18.82	19.26	18.51	21.10	20.31	18.03	20.79	20.66	16.47	19.31
	SSIM↑	0.647	0.651	0.648	0.710	0.692	0.772	0.787	0.767	0.760	0.786	0.733	0.723
	LPIPS↓	0.348	0.349	0.371	0.332	0.367	0.266	0.317	0.386	0.252	0.279	0.382	0.332
Ours	PSNR↑	19.92	19.19	19.59	20.55	19.00	22.34	23.55	22.93	22.04	22.37	20.68	21.11
	SSIM↑	0.654	0.627	0.651	0.729	0.693	0.784	0.832	0.840	0.781	0.791	0.743	0.739
	LPIPS↓	0.318	0.331	0.317	0.311	0.350	0.249	0.204	0.231	0.199	0.250	0.352	0.283

Table 3. Photorealistic mapping results in stereo EuRoC MaV dataset [1]. We color code each column as **best** and **second best**. All others are obtained in our experiments. 'X' denotes experimental failure.

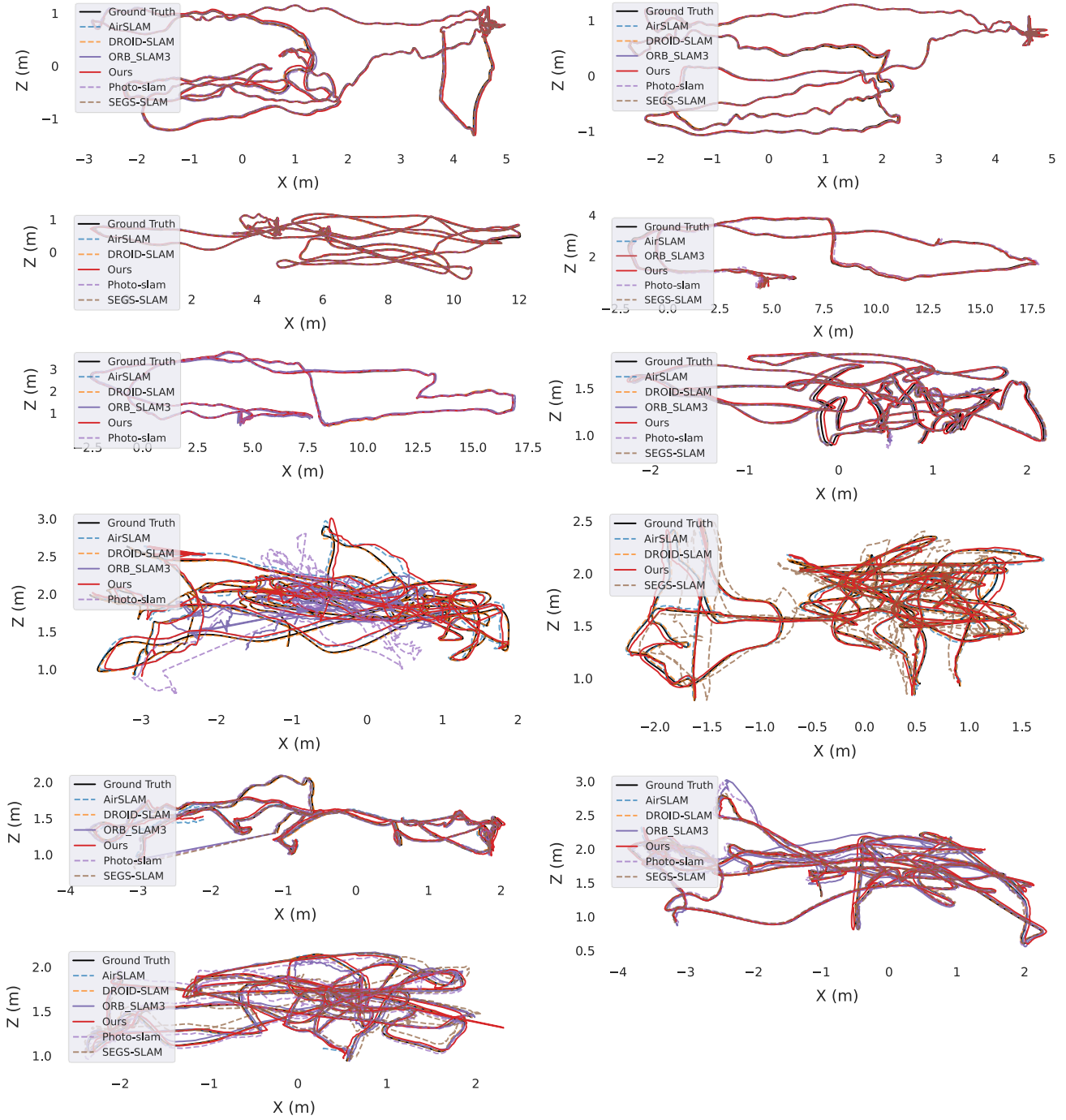


Figure 4. Per-scene qualitative results of localization in stereo EuRoC MaV dataset [1]. Specifically, from top to bottom, the left and right images in each row sequentially correspond to the sequences MH01, MH02, MH03, MH04, MH05, V101, V203, V103, V201, V202, and V102, respectively.

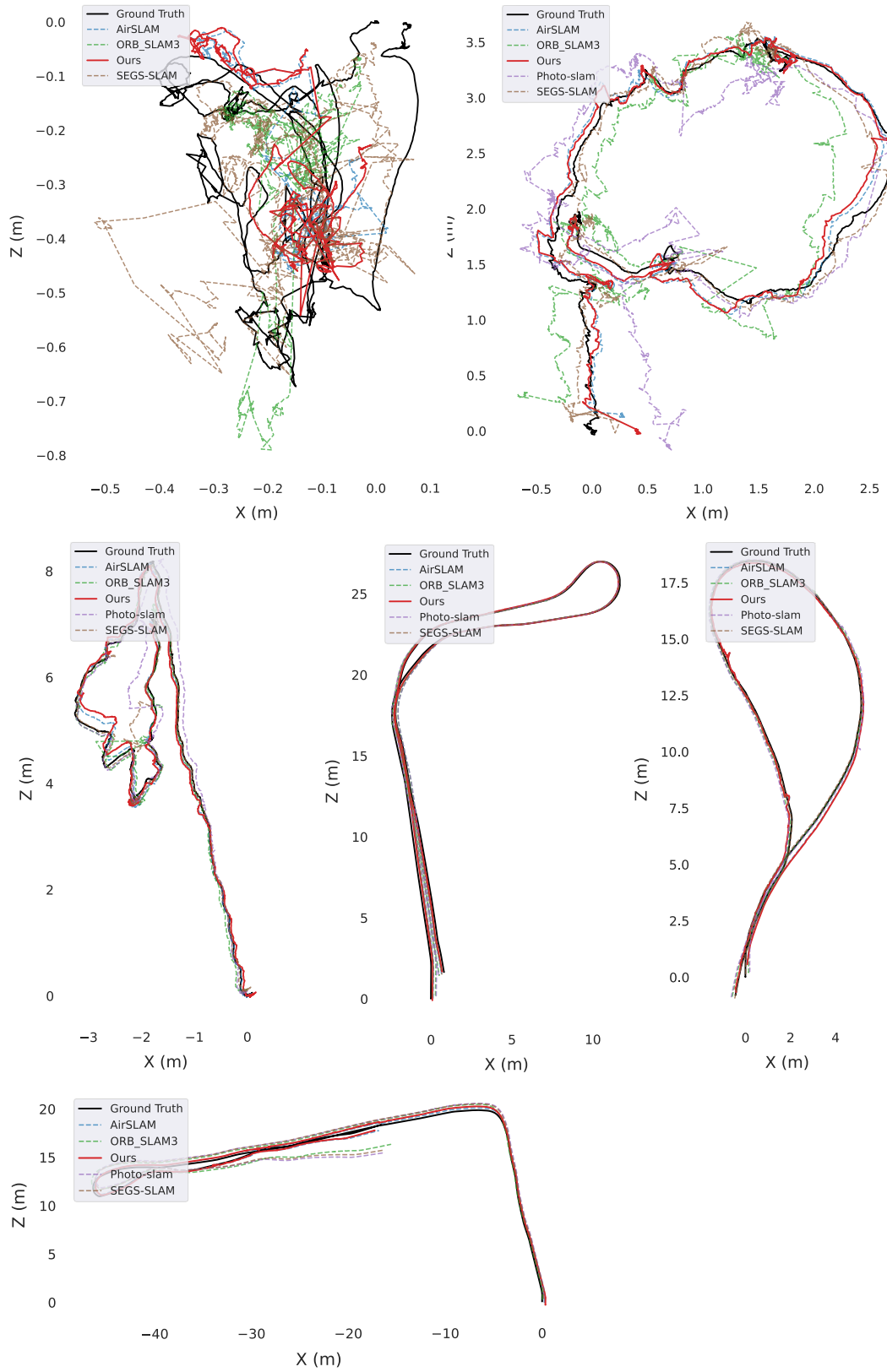


Figure 5. Per-scene qualitative results of localization in our self-collected dataset. Specifically, from top to bottom, the left and right images in each row sequentially correspond to the sequences S1, S2, S3, S4, S5, and S6, respectively.

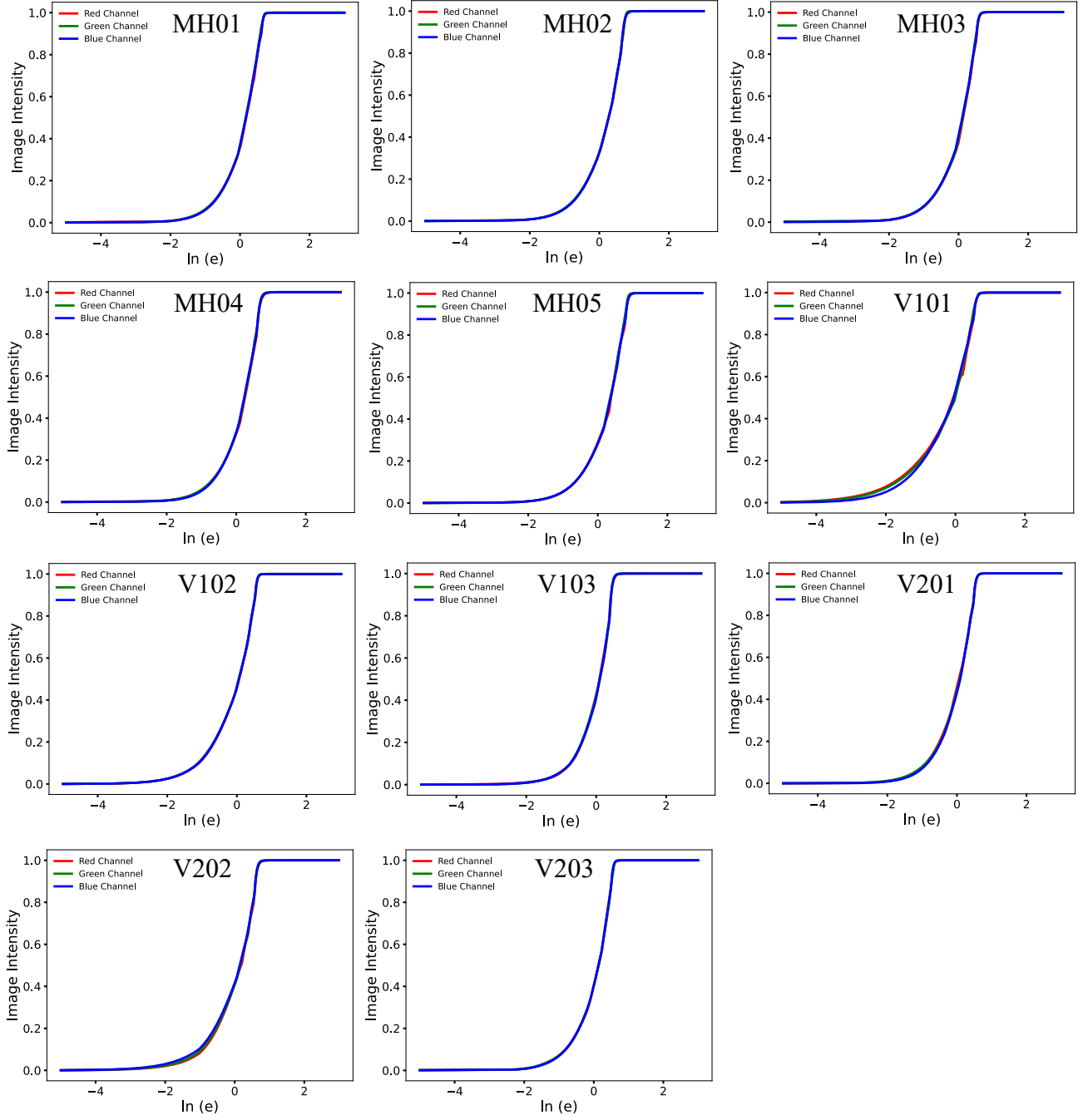


Figure 6. Per-scene CRF curves in stereo EuRoC MaV dataset [1]. Specifically, from top to bottom, the left and right images in each row sequentially correspond to the sequences MH01, MH02, MH03, MH04, MH05, V101, V102, V103, V201, V202, and V203, respectively.

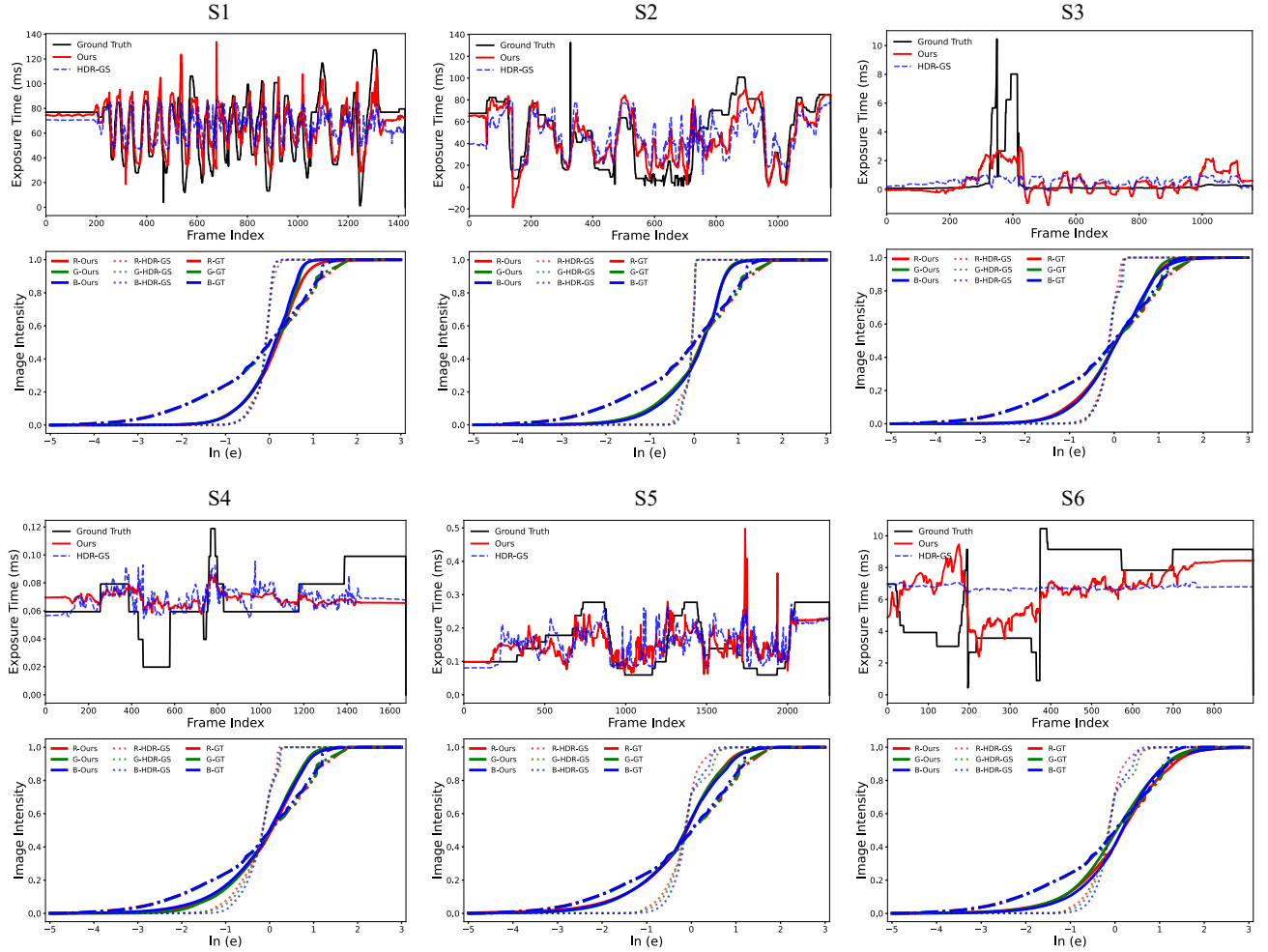


Figure 7. Per-scene CRF curves and exposure times in our self-collected dataset. Specifically, from top to bottom, the left and right images in each row sequentially correspond to the sequences S1, S2, S3, S4, S5, and S6, respectively.

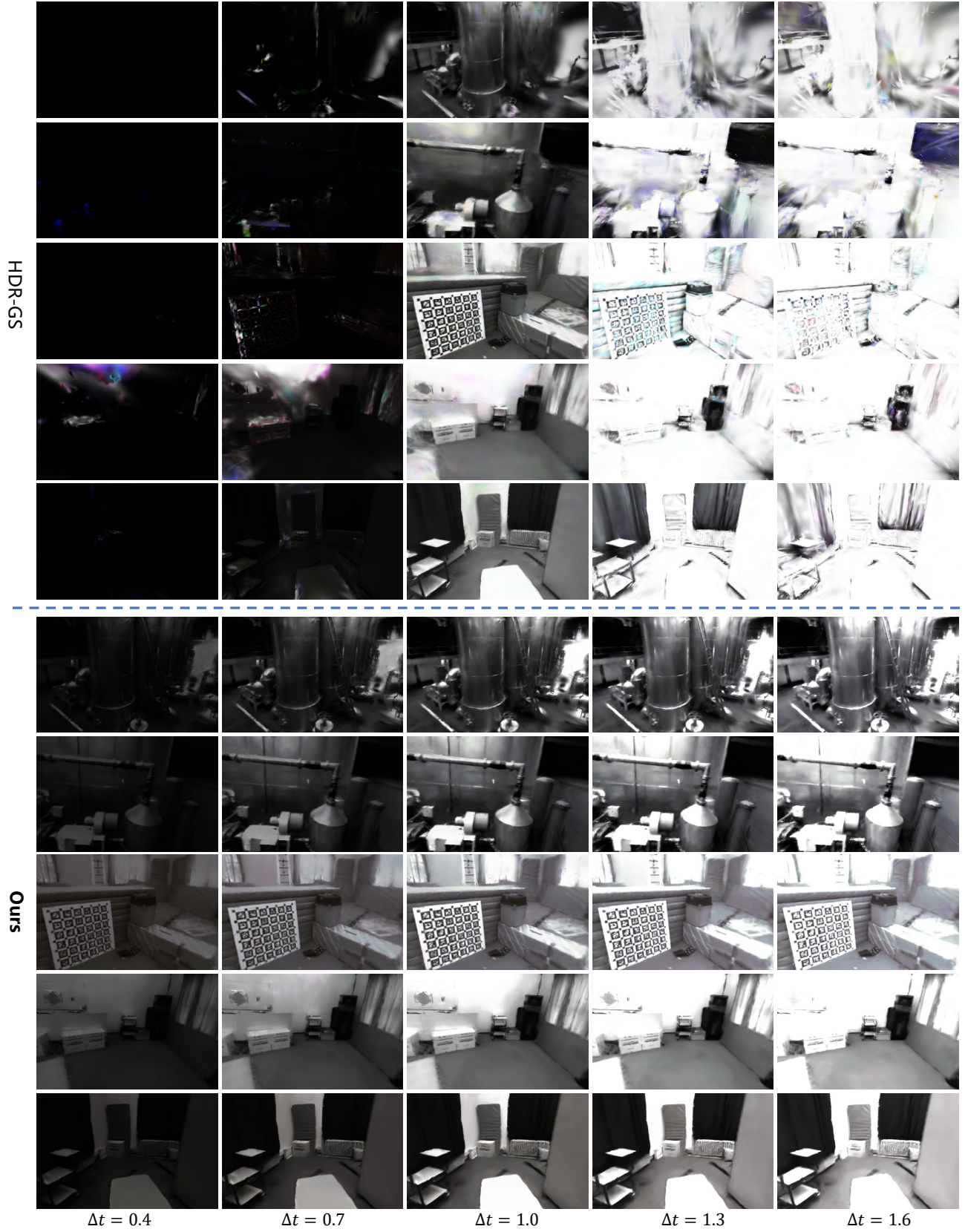


Figure 8. Exposure-controlled renderings for the EuRoC dataset [1] with exposure times Δt of 0.4, 0.7, 1.0, 1.3, and 1.6, respectively. Top row: results of HDR-GS [2]; Bottom row: results of the proposed method.

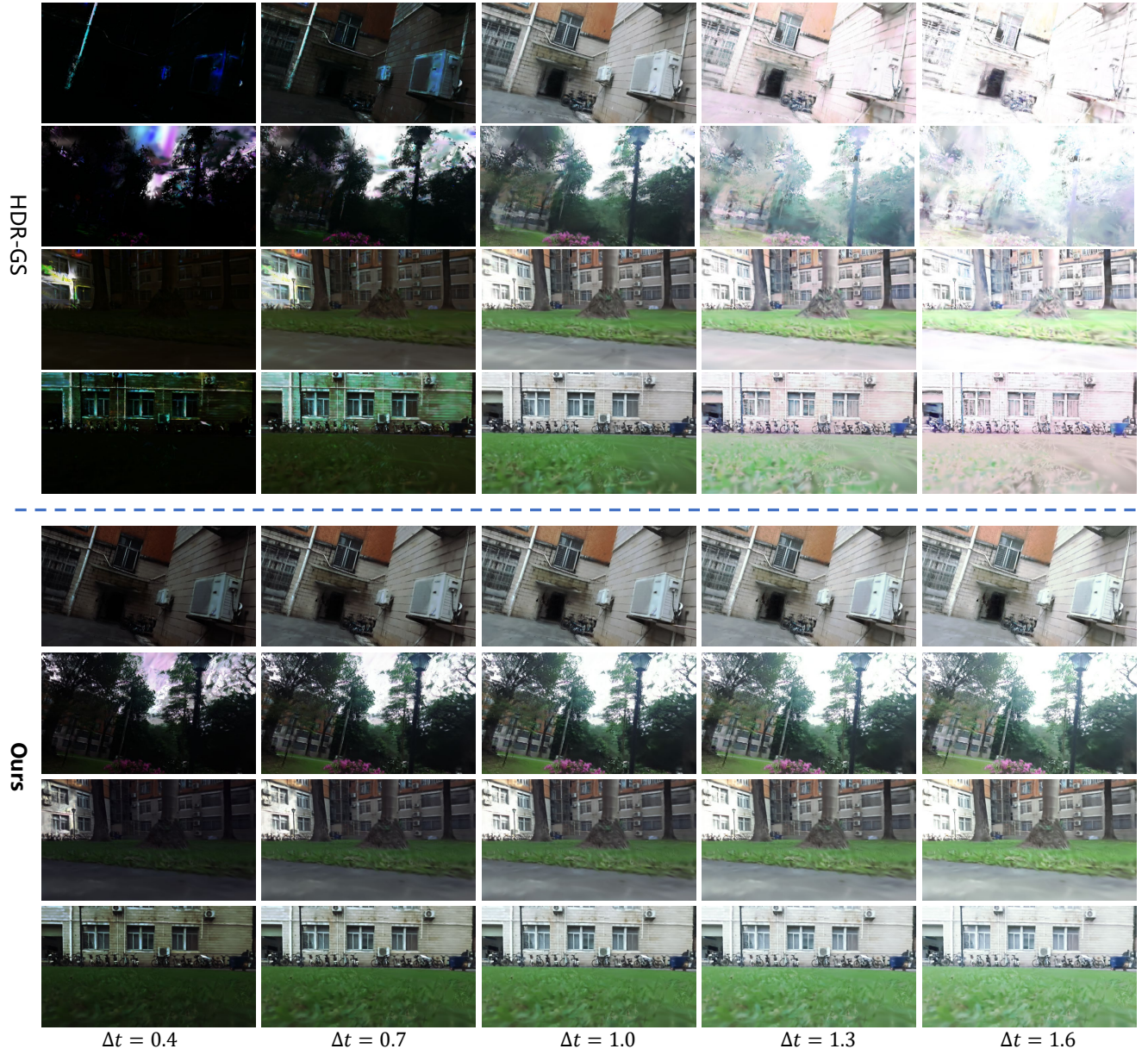


Figure 9. Exposure-controlled renderings for our self-collected dataset with exposure times Δt of 0.4, 0.7, 1.0, 1.3, and 1.6, respectively. Top row: results of HDR-GS [2]; Bottom row: results of the proposed method.



Figure 10. Qualitative comparison of diverse systems from EuRoC MAV and our self-collected dataset.

References

- [1] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, 2016. [2](#), [3](#), [4](#), [5](#), [7](#), [9](#)
- [2] Yuanhao Cai, Zihao Xiao, Yixun Liang, Minghan Qin, Yulun Zhang, Xiaokang Yang, Yaoyao Liu, and Alan Yuille. Hdr-gs: Efficient high dynamic range novel view synthesis at 1000x speed via gaussian splatting. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 68453–68471, 2024. [1](#), [2](#), [4](#), [9](#), [10](#), [11](#)
- [3] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. [3](#)
- [4] Huajian Huang, Longwei Li, Hui Cheng, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photo-realistic mapping for monocular, stereo, and rgb-d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21584–21593, 2024. [1](#), [2](#), [3](#), [4](#), [11](#)
- [5] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18039–18048, 2024. [3](#), [4](#)
- [6] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 16558–16569, 2021. [3](#)
- [7] Tianci Wen, Zhiang Liu, and Yongchun Fang. Segs-slam: Structure-enhanced 3d gaussian splatting slam with appearance embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. [1](#), [2](#), [3](#), [4](#), [11](#)
- [8] Kuan Xu, Yuefan Hao, Shenghai Yuan, Chen Wang, and Lihua Xie. Airlam: An efficient and illumination-robust point-line visual slam system. *IEEE Transactions on Robotics*, 41:1673–1692, 2025. [3](#)