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

Zhiyu Zhou Feng Hui Yu Liu *
South China University of Technology, China
{auzyzhou, auhuifeng}@mail.scut.edu.cn, auylau@scut.edu.cn

Abstract

3D Gaussian splatting (3DGS) has emerged as a revolutionary scene representation in simultaneous localization and mapping (SLAM) research. However, existing research on 3DGS-based SLAM fails to accurately address the appearance variations induced by camera auto-exposure in prevalent real-world scenarios, resulting in reduced localization and photorealistic mapping accuracy. To address this issue, we propose a stereo auto-exposure-robust Gaussian splatting SLAM (AERGS-SLAM), a framework robust to such variations and enables both reliable localization and exposure-controlled photorealistic mapping. Our key contributions are two fold. Firstly, we propose a camera exposure network to model the camera exposure process, which we integrate with Gaussian splatting to achieve exposure-controlled novel view synthesis. Secondly, we exploit an illumination-robust geometric feature for localization and Gaussian map initialization, enhancing localization accuracy under exposure-varying scenarios. Extensive experiments on public datasets and our self-collected real-world dataset demonstrate that AERGS-SLAM outperforms baselines in both localization performance and photorealistic mapping quality. Our code is available at: https://github.com/zzy-2021/AERGS-SLAM.

1. Introduction

Visual simultaneous localization and mapping (SLAM) is a fundamental problem in 3D computer vision and robotics, with wide practical applications in visual navigation, autonomous driving and other related fields. Recently, neural radiance fields (NeRF) [27] and 3D Gaussian splatting (3DGS) [20] have revolutionized 3D scene representation through photorealistic view synthesis, a capability essential for high-fidelity reconstruction. Thus, NeRF and 3DGS have been incorporated into numerous visual SLAM systems for appearance reconstruction [14, 26, 33, 37, 40, 47]. Crucially, 3DGS-based SLAM [14, 26, 37, 41, 43] and its applications [4–6, 45] have drawn substantial research interest owing to superior rendering performance.

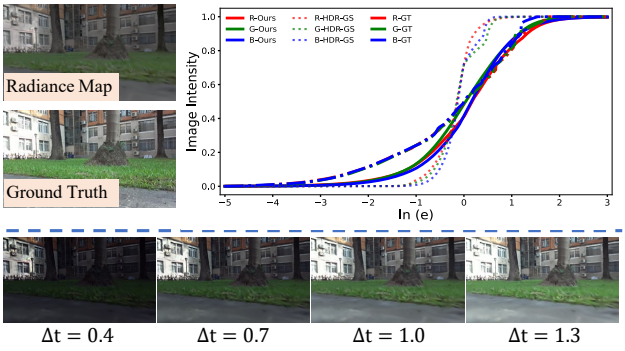


Figure 1. Estimated CRF and exposure-controlled renderings. Top row: recovered CRF curves (i.e., ours and HDR-GS [2]) and rendered scene radiance map. Bottom row: novel views generated by operating the radiance map via the camera exposure network under varying exposure time Δt.

Most 3DGS-based visual SLAM methods assume that input images strictly satisfy photometric consistency. However, to capture high-quality images, cameras automatically regulate light input via auto-exposure (AE) algorithms, which induces appearance variations in images and leads to photometric inconsistency. A few SLAM methods [26, 37] introduce per-image embeddings to address such appearance variations. For instance, MonoGS [26] adjusts image brightness via two exposure parameters, yet it fails to model complex AE mechanisms. SEGS-SLAM [37] leverages Photo-SLAM [14] and introduces view-dependent appearance embeddings to handle appearance variations. Nevertheless, AE-induced appearance variations are inherently view-independent, as they originate from camera's exposure control mechanisms, rendering them challenging to mitigate via appearance embeddings. To tackle this issue, HDR-NeRF [15] and HDR-GS [2] model the exposure process via the camera response function (CRF), which maps per-point or per-Gaussian radiance to scene color, and subsequent rendering operation is performed to synthesize images. However, such methods suffer from a key limitation:

*Corresponding author: Yu Liu

scene radiance-to-color mapping inherently couples camera exposure with the rendering process, which not only impairs appearance reconstruction quality but also increases computational overhead significantly.

Another partially unresolved challenge in 3DGS-based visual SLAM is illumination-robust localization. Coupled methods [19, 26, 39], represented by MonoGS [26], achieve localization via a 3D Gaussian map and trainable appearance parameters. However, such coupled methods suffer from key limitations in localization robustness and real-time performance. In contrast, decoupled methods [12, 14, 29, 37, 44], represented by Photo-SLAM [14], leverage traditional SLAM systems (i.e., ORB-SLAM3 [3]) for localization to ensure real-time localization. However, traditional handcrafted feature-based SLAM system lacks robustness to AE-induced illumination variations, leading to reduced localization accuracy and degraded appearance reconstruction quality in exposure-varying scenarios. Additionally, these methods only adopt multi-scale frequency representations to accelerate training, failing to account for the temporal dynamics of consecutive frames.

To address these problems, we propose a stereo decoupled auto-exposure-robust Gaussian splatting SLAM (AERGS-SLAM). Firstly, we observe that the AE process can be modeled via the CRF. While existing methods [2, 15] leverage the CRF for per-point or per-Gaussian radiance-to-color mapping, this introduces degraded rendering quality and significant computational overhead. Motivated by the physical image formation process [8], we propose a camera exposure network that models the CRF to map per-image radiance maps to red-green-blue (RGB) images, enabling exposure-controlled rendering. As shown in Fig. 1, AERGS-SLAM supports rendering RGB images and scene radiance maps. Compared with HDR-GS [2], it recovers the camera's CRF curve more accurately for exposure-controlled rendering. This CRF can be directly used to map per-image radiance maps to exposure-controlled RGB images. We adopt illumination-robust geometric features from learning-based visual SLAM [38] for localization and geometry mapping, thereby enhancing localization robustness in decoupled 3DGS-based SLAM systems. We further introduce a time-aware sliding window coarse-to-fine strategy to incorporate temporal dynamics. To summarize, the main contributions of this work are as follows:

- We propose a camera exposure network that recovers the camera's CRF to map per-image radiance maps to RGB images, enabling efficient exposure-controlled Gaussian splatting.
- We propose AERGS-SLAM, the first decoupled 3DGS-based SLAM framework leveraging a learning-based illumination-robust system to improve localization and photorealistic mapping under exposure variations.
- We conduct comprehensive experiments on public and self-collected real-world datasets, showing AERGS-SLAM outperforms baselines in mapping and localization under exposure-varying scenarios.

2. Related Work

Traditional Visual SLAM. Traditional visual SLAM falls into two categories: direct methods and indirect methods. Direct methods [9, 10] leverage photometric errors to optimize camera poses and reconstruct sparse or semi-dense geometric maps, offering high speed but susceptibility to image appearance variations. Indirect methods [3, 16, 25, 30] rely on local features (i.e., keypoints) to construct reprojection errors and reconstruct sparse geometric maps, balancing computational efficiency and accuracy. However, handcrafted features [31] show limited robustness to illumination variations. Recently, several methods [11, 23, 34, 36, 38] integrate learning-based components into traditional visual SLAM to enhance performance. DROID-SLAM [34] develops a neural network-based SLAM system to boost robustness and accuracy, yet incurs computational overhead. AirSLAM [38] adopts learning-based features to improve robustness under illumination variations. However, these methods focus on geometric mapping while neglecting the equally crucial appearance mapping.

NeRF-based SLAM. NeRF [27] implicitly encodes scene geometry and appearance via neural networks. iMAP [33] pioneers an MLP-only visual SLAM framework. Building on iMAP, numerous studies [7, 13, 18, 21, 24, 32, 40, 42, 46, 47] incorporate MLPs with other scene representations (i.e., voxel grids, triplanes, neural point clouds) to improve efficiency. More recently, NeRF-based SLAM has been extended to diverse applications [17, 22, 28, 35]. Despite this progress, these methods still overlook the impact of the exposure-induced image appearance variations on photorealistic mapping, a prevalent real-world challenge. Additionally, the volumetric rendering incurs substantial computational overhead, limiting its real-time applicability.

3DGS-based SLAM. 3DGS [20] employs anisotropic Gaussian ellipsoids to explicitly model scene geometry and appearance. 3DGS-based SLAM falls into two categories: coupled and decoupled methods. Coupled methods [19, 26, 39] adopt a unified framework for camera localization and photorealistic mapping, boosting accuracy while limiting real-time performance and robustness. In contrast, decoupled methods [14, 29, 37, 44] handle camera localization and photorealistic mapping independently, enabling real-time operation. For instance, Photo-SLAM [14] employs two parallel threads, i.e., one leveraging ORB-SLAM3 [3] for localization and other adopting 3DGS for mapping, thus ensuring real-time performance. Beyond this, other decoupled methods [29, 37, 44] also adopt similar frameworks to enhance performance, achieving a balance between accuracy and efficiency. However, none of

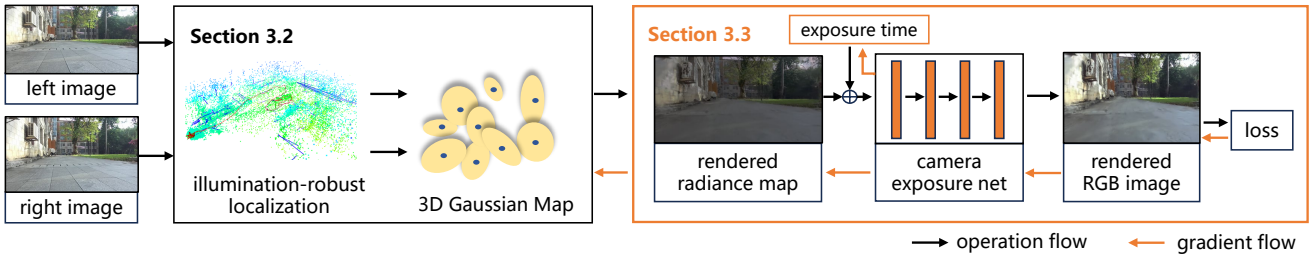


Figure 2. Overview of the proposed AERGS-SLAM. Firstly, the localization thread performs illumination-robust localization using stereo images, generating posed keyframes and sparse point clouds to initialize the Gaussian map. Secondly, the photorealistic mapping thread renders the radiance map. Subsequently, this radiance map is fed into the camera exposure network to generate RGB images and compute the loss. Finally, the resulting loss is used to optimize the Gaussian map, the camera exposure network, and the exposure time.

these methods address exposure-induced image appearance variations, which compromises photometric consistency essential to 3DGS. To fill this gap, our proposed decoupled AERGS-SLAM models the camera's AE process, enabling accurate appearance reconstruction.

3. AERGS-SLAM

The system overview of AERGS-SLAM is illustrated in Fig. 2, including a 3DGS representation map (Section 3.1), an illumination-robust localization module (Section 3.2), an exposure-controlled photorealistic mapping module (Section 3.3) and a loop closure detection module (Section 3.4).

3.1. 3DGS Representation

For the 3DGS radiance field representation, we employ anisotropic Gaussian ellipsoids [20] to model the scene's geometry and appearance. A Gaussian ellipsoid is parameterized by its position mean $\mathbf{P} \in \mathbb{R}^3$, covariance $\mathbf{\Sigma} \in \mathbb{R}^6$, spherical harmonics $\mathbf{S} \in \mathbb{R}^k$ (k is the degrees of freedom) and opacity α. Spherical harmonics $\mathbf{S}$ is used to recover radiance $\mathbf{e} \in \mathbb{R}^3$. Mathematically, a Gaussian ellipsoid follows a Gaussian distribution

$$G(\mathbf{X}) = e^{-\frac{1}{2}\mathbf{X}^\top \mathbf{\Sigma}^{-1}\mathbf{X}}, \tag{1}$$

where $G(\mathbf{X})$ denotes the density of the Gaussian ellipsoid at position $\mathbf{X}$. Following 3DGS [20], for a given view frustum, 3DGS queries N Gaussians within the frustum, forming a Gaussian set $\mathcal{G} = \{G_1, \ldots, G_N\}$. This set $\mathcal{G}$ is employed to synthesize photorealistic images via two steps: 1) projecting the Gaussian set $\mathcal{G}$ onto the image plane to form the corresponding 2D Gaussian set $\mathcal{G}' = \{G'_1, \ldots, G'_N\}$ and 2) sorting the 2D Gaussian set $\mathcal{G}'$ from farthest to nearest and applying the α-blending

$$\mathbf{I} = \sum_{i \in N} \mathbf{e}_i \cdot G'_i \alpha_i \prod_{j=1}^{i-1} (1 - G'_j \alpha_j), \tag{2}$$

where G'_i is the i-th sorted 2D Gaussian ellipsoid; $\alpha_i (i = 1, \ldots, N)$ is the opacity of G'_i; $\mathbf{e}_i (i = 1, \ldots, N)$ is the radiance of G'_i; and $\mathbf{I}$ is the rendered image. Both equations (1) and (2) are differentiable, allowing the Gaussians $\mathcal{G}$ to be optimized via photometric loss.

3.2. Illumination-Robust Localization

The localization module utilizes a visual SLAM pipeline [38] to efficiently estimate camera poses and generate sparse 3D landmarks. A camera pose is represented by rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. We adopt a sliding window-based method to jointly estimate camera poses and landmark positions.

Given a sliding window with K keyframes, we define a keyframe set $\mathcal{F} = \{F_1, \cdots, F_K\}$, a rotation set $\mathcal{R} = \{\mathbf{R}_1, \cdots, \mathbf{R}_K\}$, and a translation set $\mathcal{T} = \{\mathbf{t}_1, \cdots, \mathbf{t}_K\}$. For each keyframe F_k $(k = 1, \cdots, K)$, m_k denotes the number of 2D-3D matches in F_k and $\mathcal{X}^k = \{X_{k1}, \cdots, X_{km_k}\}$ denotes the 2D-3D match set, where X_{kj} is the j-th match in F_k. All 2D-3D matches in the window are $\mathcal{X} = \{\mathcal{X}^1, \cdots, \mathcal{X}^K\}$. Camera rotations $\mathcal{R}$, translations $\mathcal{T}$ and 3D landmarks associated with $\mathcal{X}$ are solved via local bundle adjustment (BA), given by

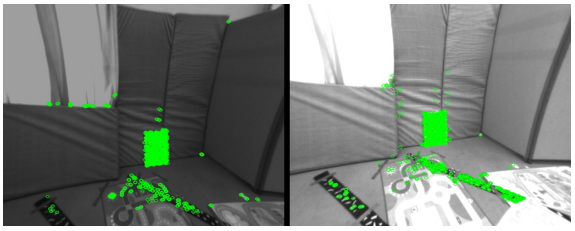
$$\mathcal{R}, \mathcal{T}, \mathcal{X} = \arg\min_{\mathcal{R},\mathcal{T},\mathcal{X}} \sum_{k=1}^{K} \sum_{j=1}^{m_k} \rho(E(k, j)), \tag{3}$$

where $\rho(\cdot)$ is a robust kernel function; $E(k, j)$ is the residual of the j-th match X_{kj} in F_k, defined as

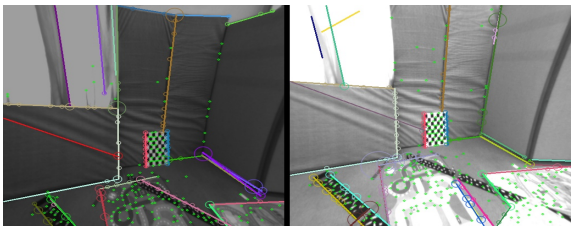
$$E(k, j) = \|\mathbf{p}_{kj} - \pi(\mathbf{R}_k \mathbf{P}_{kj} + \mathbf{t}_k)\|^2, \tag{4}$$

where $\mathbf{p}_{kj}$ is the 2D pixel coordinate of X_{kj} in F_k; $\mathbf{P}_{kj}$ is the corresponding 3D landmark coordinate in the world frame; $\pi(\cdot)$ denotes the camera projection function.

Equations (3) and (4) show that the feature matching accuracy directly determines the reliability of residuals $E(k, j)$. Inaccurate matches degrade $E(k, j)$, undermining BA's pose and landmark estimation, while precise matches ensure robust optimization. Specifically, camera exposure variations induce image appearance changes, and

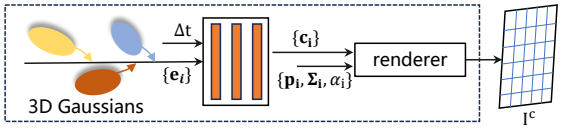


(a) Handcrafted ORB features [31].

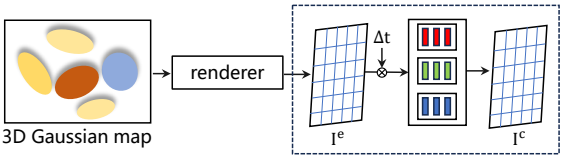


(b) Learning-based illumination-robust features [38].

Figure 3. Feature detection in illumination-varying scene.



(a) Exposure modeling in HDR-GS [2].



(b) The proposed camera exposure network.

Figure 4. Exposure modeling in HDR-GS [2] and the proposed method. The differences between them are: 1) HDR-GS uses a network to map per-Gaussian radiance $\mathbf{e}_i$ to per-Gaussian color $\mathbf{c}_i$, while CEN maps the radiance map $\mathbf{I}^e$ to color map $\mathbf{I}^c$. 2) CEN decouples the rendering and exposure processes to achieve rendering-independent exposure estimation.

handcrafted features [3] lack sufficient robustness to such appearance variations, reducing the reliability of residual $E(k, j)$. As shown in Fig. 3a, handcrafted ORB [3] features are not robust to AE-induced image appearance variations. By contrast, Fig. 3b adopts learning-based features [38] for illumination-robust feature extraction and matching. By incorporating illumination-robust features, the BA problem can be solved more accurately under exposure-varying scenarios. In our AERGS-SLAM system, the optimized keyframe set $\mathcal{F}$ and 3D landmark set $\{\mathbf{P}_{kj}\}$ serve as training views and initial Gaussian ellipsoids for photometric mapping, respectively.

3.3. Auto-Exposure-Robust Mapping

3.3.1. Camera Exposure Network

A key limitation of 3DGS is its reliance on multi-view photometric consistency, yet the camera's AE algorithm induces image appearance variation, undermining this consistency. Exposure-controlled radiance fields [2, 15] have been verified as a solution to this problem. However, as shown in Fig. 4a, HDR-GS [2] maps per-Gaussian radiance $\{\mathbf{e}_i\}$ to color $\{\mathbf{c}_i\}$, leading to coupling between the camera exposure and the rendering process. This coupling reduces the appearance modeling capability and increases computational overhead. To address this, we propose a camera exposure network that models the CRF to map the per-image radiance map $\mathbf{I}^e$ to the RGB image $\mathbf{I}^c$, enables more accurate exposure-controlled rendering.

As shown in Fig. 4b, we decouple the rendering and exposure process. The 3DGS map renders the radiance map $\mathbf{I}^e$ and CEN maps $\mathbf{I}^e\Delta t$ to the RGB image $\mathbf{I}^c$. This is mathematically described by

$$\mathbf{I}^c = f(\mathbf{I}^e \Delta t), \tag{5}$$

where $f(\cdot)$ denotes CRF. Following the CRF calibration method by Debevec and Malik [8], we transform the radiance map into the logarithmic radiance domain. Assuming the CRF f is monotonic and invertible, so we rewrite (5) as

$$\ln f^{-1}(\mathbf{I}^c) = \ln(\mathbf{I}^e) + \ln(\Delta t). \tag{6}$$

We denote the inverse function of $\ln f^{-1}(\cdot)$ as $g(\cdot)$, thus

$$\mathbf{I}^c = g(\ln \mathbf{I}^e + \ln \Delta t), \tag{7}$$

where $g(\cdot) = (\ln f^{-1}(\cdot))^{-1}$ and is estimated via a network. In our work, we employ three independent MLPs to model the RGB channels respectively. Combined with (2), the exposure-controlled rendering equation is derived as

$$\mathbf{I}^c = g(\ln \sum_{i \in N} \mathbf{e}_i \cdot G'_i \alpha_i \prod_{j=1}^{i-1} (1 - G'_j \alpha_j) + \ln \Delta t). \tag{8}$$

Equation (8) shows that exposure control and radiance map rendering are computed independently. Compared with HDR-GS [2], the proposed CEN offers two key advantages: 1) We use the rendered radiance map for exposure adjustment or estimation, eliminating the need to recompute the rendering equation; 2) The complexity of our method depends solely on the network architecture and rendered image size, rather than the number of Gaussian ellipsoids, thus reducing algorithmic complexity. As shown in Fig. 5, the proposed CEN reconstructs the scene appearance more accurately than HDR-GS [2] without color distortion.

The photometric loss $\mathcal{L}_c$ is used to optimize Gaussian parameters $\mathcal{G}$, MLPs $g(\cdot)$ and exposure time Δt, given by

$$\mathcal{L}_c = (1 - \lambda)|\mathbf{I}^c - \mathbf{I}^c_{\text{gt}}|_1 + \lambda(1 - \text{SSIM}(\mathbf{I}^c, \mathbf{I}^c_{\text{gt}})), \tag{9}$$

Figure 5. Exposure-controlled RGB renderings of CEN and HDR-GS [2] under varying exposure times Δt.

where $\text{SSIM}(\mathbf{I}^c, \mathbf{I}^c_{gt})$ denotes structural similarity between the two images and λ is a weight factor.

We adopt the unit exposure loss $\mathcal{L}_u = \|g(0) - C_0\|_2^2$ from [15], where C_0 is set the midway of the pixel value. This loss fixes the scale of the radiance map. The final loss for photorealistic mapping is

$$\mathcal{L} = \mathcal{L}_c + \lambda_u \mathcal{L}_u, \tag{10}$$

where λ_u denotes the weight of unit exposure loss.

3.3.2. Coarse-To-Fine Optimization

Coarse-to-fine optimization strategy is effective in many SLAM methods. For example, methods [40, 47] use hierarchical components for multi-level details. Methods [14, 37] adopt multi-scale frequency representations to accelerate training. However, these methods use a fixed low-to-high frequency progression for the entire scene and overlook the temporal dynamics of SLAM keyframes, where new and old keyframes inherently carry distinct temporal information. To mitigate this limitation, we propose a time-aware sliding window coarse-to-fine strategy. New keyframes with incompletely reconstructed scene information supervise low-frequency structures, while old keyframes supervise high-frequency appearance details, with the optimization proceeding dynamically within the sliding window.

To implement this, we design a novel image sampling strategy within the sliding window. Specifically, we define a sliding window containing L keyframes, ordered from earliest to most recently observed and denoted as $\{F_l\}(l = 1, \cdots, L)$. Each keyframe is assigned a residence time $N_l(l = 1, \cdots, L)$, representing its duration in the window. We introduce a scaling function $h(\cdot)$ to compute the downsampling scale $\alpha_l = h(N_l)$ for the l-th keyframe F_l. Older keyframes with longer N_l are assigned smaller α_l to retain high-frequency details, while newer keyframes use larger α_l for low-frequency supervision. The optimization over the sliding window $\{F_l\}$ is formulated as

$$\begin{aligned} F_1 &: \arg\min \mathcal{L}(\mathbf{I}^1_r, \text{sample}(\mathbf{I}^1_{\text{gt}}, \alpha_1)), \\ &\cdots \\ F_L &: \arg\min \mathcal{L}(\mathbf{I}^L_r, \text{sample}(\mathbf{I}^L_{\text{gt}}, \alpha_L)), \end{aligned} \tag{11}$$

where $\mathbf{I}^l_{\text{gt}}$ is the ground truth of the l-th keyframe; $\mathbf{I}^l_r$ is the rendering image of the l-th keyframe; $\text{sample}(\cdot)$ denotes the downsampling function with a scale factor α_l; and $\mathcal{L}$ corresponds to (10). Experiments validate that this sliding window-based training strategy enhances photorealistic mapping performance.

3.4. Loop Closure

Loop closure [38] detects keyframes that are temporally separated but spatially close, and establishes new 2D-3D constraints between these keyframes. These detected loop keyframes and new constraints are used to formulate the BA problem defined in equation (4) to correct camera poses and reduce trajectory drift. With the corrected camera poses, the photorealistic mapping module leverages these poses to perform appearance reconstruction and exposure estimation, thereby further improving the accuracy of photorealistic mapping in AERGS-SLAM.

4. Experiments

In our experiments, Section 4.1 shows the implementation details. Section 4.2 describes the experiment setup. Section 4.3 reports the experimental results and evaluation. Section 4.4 reports ablation studies. Notably, for more comprehensive qualitative and quantitative results, we **strongly** encourage readers to refer to the **supplementary material**.

4.1. Implementation Details

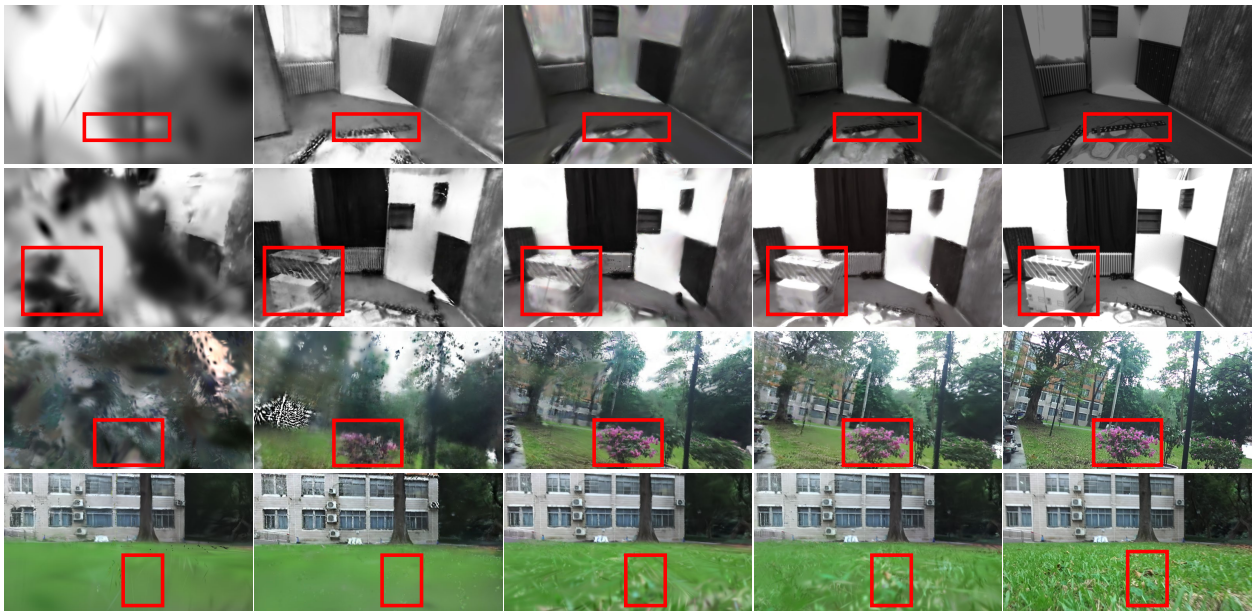
The proposed AERGS-SLAM is fully implemented in C++ and the LibTorch framework. The localization and photorealistic mapping modules operate in parallel. The localization module generates sparse point clouds and posed keyframes. The localization module generates sparse point clouds and posed keyframes, which are fed to the mapping module to initialize Gaussian ellipsoids and train the 3D Gaussian map. These keyframes act as the training set, with remaining frames used as the testing set. The localization module adopts the default settings from AirSLAM [38]. The learning rate for Gaussian parameters follows Photo-SLAM [14]. The learning rate for the MLP and exposure time are set to 0.001 and 0.02, respectively. We set λ = 0.4, λ_u = 0.5. For sliding window optimization, the scaling function is $h(N_l) = -0.065N_l + 8$ for $N_l \leq 100$, and fixed at 1.5 when $N_l > 100$. The midway of the pixel value C_0 is set to 0.73.

4.2. Experiment Setup

Baselines. We compare AERGS-SLAM with seven baselines: 1) MonoGS [26], a state-of-the-art (SOTA) coupled 3DGS-based SLAM method; 2) Photo-SLAM [14] and SEGS-SLAM [37], representative decoupled 3DGS-based methods; 3) Ours + HDR-GS, a variant where our CEN is replaced with HDR-GS's exposure modeling [2] to validate

Method	EuRoC MAV						Self-collected					
	MH01	MH03	V102	V103	V202	V203	S1	S2	S3	S4	S5	S6
ORB-SLAM3 [3]	0.044	X	0.088	X	0.125	1.522	0.359	0.523	0.643	2.481	3.423	1.716
DROID-SLAM [34]	0.012	0.022	0.012	0.019	0.010	0.055	-	-	-	-	-	-
AirSLAM [38]	0.022	0.023	0.020	0.031	0.022	0.218	0.192	0.155	0.095	0.168	0.478	0.234
MonoGS [26]	0.089	1.821	0.042	0.745	1.592	X	X	3.479	2.587	16.975	32.227	9.265
Photo-SLAM [14]	0.029	0.035	0.080	X	0.064	1.001	X	0.366	0.627	2.500	3.553	1.723
SEGS-SLAM [37]	0.037	0.052	0.161	0.288	0.062	X	0.211	0.426	0.637	2.473	3.633	1.680
Ours	0.021	0.023	0.051	0.024	0.023	0.215	0.177	0.167	0.137	0.132	0.523	0.264

Table 1. Quantitative results of localization (RMSE ↓). We color code eac column as best and second best. 'X' denotes running failure in our experiments. '-' denotes no results, as we use DROID-SLAM's results as reference poses on the self-collected dataset.



(a) Photo-SLAM [14] (b) SEGS-SLAM [37] (c) Ours-HDRGS [2] (d) **Ours** (e) Ground Truth

Figure 6. Qualitative comparison of diverse systems from EuRoC MAV and our self-collected dataset.

CEN's effectiveness; 4) ORB-SLAM3 [3], a classic hand-crafted feature-based SLAM method; 5) DROID-SLAM [34] and AIR-SLAM [38], learning-enhanced methods.

Hardware. We run AERGS-SLAM and all baseline methods using their official implementations on a desktop computer equipped with an RTX 4090 24GB GPU, an Intel Core i7-13900K CPU, and 32 GB of RAM.

Dataset. First of all, to our knowledge, no public SLAM dataset explicitly evaluates AE robustness. Inspired by [38], we process the EuRoC MAV dataset [1] by adjusting image brightness to simulate AE-induced exposure variations. Brightness adjustment is modeled as $V_{\text{out}} = AV_{\text{int}}$, where V_{int} and V_{out} are the input and output brightness of a pixel, respectively, and A is the scaling factor which is randomly sampled uniformly within $[0.5, 1.5]$ for each image. All baselines are evaluated in this processed dataset. Then, we evaluate on our self-collected dataset, which consists of six sequences captured using a ZED 2i stereo camera. For quantitative evaluation, we adopt the trajectory from the SOTA learning-based stereo SLAM system DROID-SLAM [34] as the reference. Given its demonstrated superior performance in handling complex real-world scenarios and stereo setups in recent literature [14, 37], DROID-SLAM provides a reliable benchmark for assessing our method's localization accuracy. Additionally, we recorded real exposure times to evaluate exposure estimation performance.

Metrics. For localization, we report the root mean square error (RMSE) of the absolute trajectory error for all frames. For photorealistic mapping, we report the PSNR, SSIM, and LPIPS metrics to evaluate the quality of novel view synthesis. For exposure estimation, we report the estimated relative exposure time.

Method	Metric	EuRoC MAV						Self-collected					
		MH01	MH03	V102	V103	V202	V203	S1	S2	S3	S4	S5	S6
MonoGS [26]	PSNR↑	19.07	14.99	15.32	14.90	12.86	X	X	19.28	16.81	19.48	16.99	17.70
	SSIM↑	0.757	0.587	0.750	0.730	0.634	X	X	0.772	0.501	0.555	0.492	0.424
	LPIPS↓	0.255	0.475	0.472	0.569	0.632	X	X	0.502	0.560	0.558	0.648	0.538
Photo-SLAM [14]	PSNR↑	12.02	11.62	11.34	X	11.06	10.17	X	18.47	17.09	19.97	16.04	20.03
	SSIM↑	0.341	0.386	0.577	X	0.541	0.547	X	0.744	0.489	0.551	0.486	0.481
	LPIPS↓	0.465	0.542	0.588	X	0.560	0.592	X	0.429	0.575	0.496	0.601	0.449
SEGS-SLAM [37]	PSNR↑	15.99	16.34	14.98	15.51	14.14	X	16.57	21.37	18.83	19.74	19.75	19.34
	SSIM↑	0.621	0.654	0.730	0.748	0.716	X	0.743	0.810	0.580	0.642	0.628	0.568
	LPIPS↓	0.324	0.274	0.327	0.332	0.272	X	0.379	0.294	0.388	0.446	0.415	0.442
Ours + HDR-GS [2]	PSNR↑	18.99	18.82	20.31	18.03	20.66	16.47	17.91	21.93	20.71	20.49	20.73	19.29
	SSIM↑	0.647	0.648	0.787	0.767	0.786	0.733	0.769	0.815	0.642	0.655	0.639	0.543
	LPIPS↓	0.348	0.371	0.317	0.386	0.279	0.382	0.324	0.344	0.314	0.389	0.393	0.406
Ours	PSNR↑	19.92	19.59	23.55	22.93	22.37	20.68	18.23	21.60	20.44	20.49	19.61	19.97
	SSIM↑	0.654	0.651	0.832	0.840	0.791	0.743	0.771	0.814	0.643	0.651	0.600	0.548
	LPIPS↓	0.318	0.317	0.204	0.231	0.250	0.352	0.323	0.337	0.292	0.373	0.440	0.377

Table 2. Quantitative results of Photorealistic mapping results. We color code each column as best and second best. 'X' denotes running failure in our experiments.

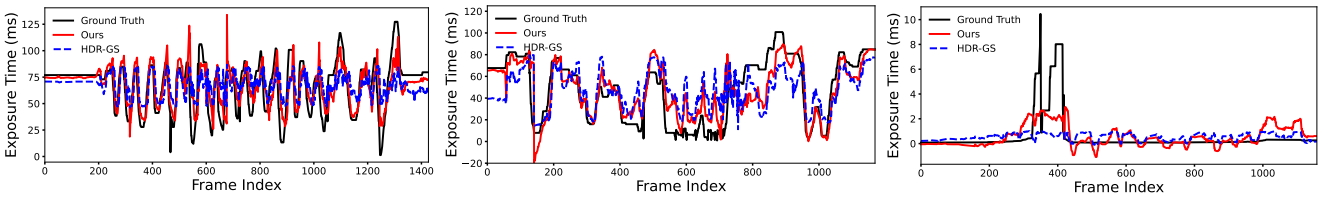


Figure 7. Camera exposure time estimation for scenes 'S1', 'S2', and 'S3'.

4.3. Results and Evaluation

Localization Accuracy. Firstly, the quantitative results are reported in Table 1. On the processed EuRoC dataset [1], AERGS-SLAM achieves the best localization performance against 3DGS-based baselines. Specifically, compared with Photo-SLAM [14] and SEGS-SLAM [37] using a handcrafted feature (i.e., ORB-SLAM3 [3]), AERGS-SLAM runs successfully on all sequences, validating the robustness of the learning-based features [38] employed in our work. Additionally, compared with MonoGS [26], all decoupled pipelines achieve superior accuracy, highlighting the robustness of the decoupled framework. Secondly, on the self-collected dataset, AERGS-SLAM achieves significantly higher localization accuracy than all baselines, confirming its generalization to real-world scenarios. Overall, these comprehensive evaluation results validate the effectiveness of our illumination-robust localization pipeline.

Novel View Synthesis. Firstly, the quantitative results of photorealistic mapping are reported in Table 2. For the EuRoC dataset [1], AERGS-SLAM outperforms Photo-SLAM [14] without using any exposure mechanism, SEGS-SLAM utilizing appearance embedding, and MonoGS utilizing learnable exposure parameters. This demonstrates CEN's ability to address AE-induced appearance variations. Moreover, compared with HDR-GS [2] mapping per-Gaussian radiance to color, our CEN mapping per-image radiance maps to RGB images, is more effective in handling camera exposure variations. Then, for our self-collected dataset, we obtain equally high-quality photorealistic mapping results in evaluations. These results further demonstrate the generalization capability of the CEN in real-world scenarios. Secondly, the qualitative results are shown in Fig. 6. We observe that AERGS-SLAM reconstructs the most realistic appearance, while other methods fail to capture fine details or adapt to exposure variations.

Exposure Estimation. Since the EuRoC MAV dataset [1] lacks ground truth exposure times, we evaluate on our self-collected dataset, where we report the estimated exposure times (in milliseconds). As shown in Fig. 7, as frame indices change, the real camera exposure time adjusts automatically, and AERGS-SLAM can synchronously estimate the camera's exposure time. Moreover, compared with HDR-GS [2], AERGS-SLAM performs better, as its esti-

Datasets	EuRoC		Self-Collected	
Method	PSNR ↑	RMSE ↓	PSNR ↑	RMSE ↓
(1) w/o CTFO, CEN, IRL	11.76	0.164	18.32	1.754
(2) w/o CTFO, CEN	14.76	0.072	19.59	0.199
(3) w/o CEN	15.10	0.051	19.69	0.214
(4) w/o CTFO	20.48	0.077	20.05	0.199
(5) Ours	21.11	0.049	20.06	0.233

Table 3. Ablation Study on the key components (1) - (5), including CTFO, CEN, and IRL. The best results are highlighted.

(a) HDR-GS [2] with rendering speed of 416 FPS.

(b) Ours with rendering speed of 3700 FPS.

Figure 8. Exposure-controlled rendering. From left to right, we set exposure time Δt to 0.5, 1.0 and 1.5.

mates are closer to the ground truth. Then, we present qualitative results of exposure-controlled rendering in Fig. 8. We observe that, compared with HDR-GS, the proposed CEN can achieve higher-quality exposure-controlled renderings. Moreover, we render nearly 10 times faster than HDR-GS. These results strongly validate CEN's effectiveness for camera exposure estimation.

4.4. Ablation Studies

Our ablation experiments validate three core modules: illumination-robust localization (IRL), coarse-to-fine optimization (CTFO), and camera exposure network (CEN). The ablation results are reported in Table 3 within Row (1) (i.e., without CTFO, CEN and IRL) corresponds to the original Photo-SLAM [14]. Experiments are conducted on the processed EuRoC and self-collected datasets. We report the average RMSE and average PSNR metrics to evaluate localization and photorealistic mapping performance.

Illumination-robust localization. As shown in the rows (1) and (2) of Table 3, when IRL is applied (i.e., the row (2) of Table 3), the RMSE metric is significantly reduced on both the EuRoC dataset and our self-collected dataset. These results strongly demonstrate the effectiveness of the learning-based features in our work. Moreover, with the improvement of localization accuracy, the PSNR metric is further enhanced. This is because higher localization accuracy can provide the mapping module with more accurate Gaussian ellipsoids and camera poses, thereby improving the accuracy of photorealistic mapping.

Coarse-to-fine Optimization. First of all, as shown in the rows (4) and (5) of Table 3, our full method (i.e., row (5)) achieves the best PSNR metric for photorealistic mapping against the method using CTFO (i.e., row (4)). Notably, the CTFO module only contributes to the photorealistic mapping module and not to the localization module. Therefore, the RMSE metrics in the rows (4) and (5) do not reflect the contribution of the CTFO module. Then, the contribution of the CTFO module can also be reflected by the change in PSNR metrics between the methods in the rows (2) and (3). These ablation results consistently demonstrate that the proposed time-aware coarse-to-fine optimization strategy can effectively improve the quality of photorealistic mapping.

Camera exposure network. Firstly, as shown in the rows (3) and (5) of Table 3, the employment of the CEN module significantly enhances the quality of photometric mapping and improves the PSNR metrics on both the EuRoC dataset and our self-collected dataset. Similarly, the CEN module only contributes to the photorealistic mapping module. Therefore, the slight changes in RMSE metrics do not reflect the contribution of the CEN module. Moreover, as shown in the rows (2) and (4) of Table 3, applying the CEN module increases the PSNR metric by more than 5 dB on the EuRoC dataset. Overall, these ablation results fully validate that the proposed CEN module is effective in boosting the quality of photorealistic mapping, especially in enhancing photometric consistency under varying exposure.

5. Conclusion

In this paper, we propose a 3DGS-based SLAM framework called AERGS-SLAM. It adopts a decoupled pipeline enabling illumination-robust localization and auto-exposure-robust photorealistic mapping. To this end, we propose a learning-based feature for IRL. Extensive experiments show the IRL module significantly improves localization accuracy and robustness. Furthermore, proposed CTFO and CEN modules effectively enhance photorealistic mapping quality under camera exposure variations. Comprehensive real-world experiments show the CEN module not only synthesizes high-fidelity novel views but also recovers per-image exposure times, enabling exposure-controlled Gaussian splatting.

6. Acknowledgment

This work was supported in part by the Special Fund for Research on National Major Research Instruments of National Natural Science Foundation of China under Grant 62527806 and in part by the Key Program of National Natural Science Foundation of China under Grant 62433011.

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. *International Journal of Robotics Research*, 35(10): 1157–1163, 2016. 6, 7

[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, 5, 6, 7, 8

[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. 2, 4, 6, 7

[4] Zhenzhong Cao, Chenyang Zhao, Qianyi Zhang, Jinzheng Guang, Yinuo Song, and Jingtai Liu. Rgbds-slam: A rgb-d semantic dense slam based on 3d multi level pyramid gaussian splatting. *IEEE Robotics and Automation Letters*, 10(5): 4778–4785, 2025. 1

[5] Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, and Yi Xu. Activegamer: Active gaussian mapping through efficient rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16486–16497, 2025.

[6] Timothy Chen, Ola Shorinwa, Joseph Bruno, Aiden Swann, Javier Yu, Weijia Zeng, Keiko Nagami, Philip Dames, and Mac Schwager. Splat-nav: Safe real-time robot navigation in gaussian splatting maps. *IEEE Transactions on Robotics*, 41:2765–2784, 2025. 1

[7] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H. Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 9400–9406, 2023. 2

[8] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the Annual Conference on Computer graphics and interactive techniques*, pages 369–378, 1997. 2, 4

[9] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *Proceedings of the European Conference on Computer Vision*, pages 834–849, 2014. 2

[10] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018. 2

[11] Taimeng Fu, Shaoshu Su, Yiren Lu, and Chen Wang. islam: Imperative slam. *IEEE Robotics and Automation Letters*, 9 (5):4607–4614, 2024. 2

[12] Seongbo Ha, Jiung Yeon, and Hyeonwoo Yu. Rgbd gs-icp slam. In *Proceedings of the European Conference on Computer Vision*, pages 180–197, 2025. 2

[13] Jiarui Hu, Mao Mao, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Cp-slam: collaborative neural point-based slam. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pages 39429–39442, 2023. 2

[14] 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, 5, 6, 7, 8

[15] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18377–18387, 2022. 1, 2, 4, 5

[16] Feng Hui, Zhiyu Zhou, and Yu Liu. Pl-lvi: A lidar-visual-inertial slam system integrating visual point-line features. *IEEE Transactions on Automation Science and Engineering*, 23:3457–3468, 2026. 2

[17] Sijia Jiang, Jing Hua, and Zhizhong Han. Query quantized neural slam. In *Proceedings of the Annual AAAI Conference on Artificial Intelligence*, pages 4057–4065, 2025. 2

[18] Mohammad Mahdi Johari, Camilla Carta, and François Fleuret. Eslam: Efficient dense slam system based on hybrid representation of signed distance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17408–17419, 2023. 2

[19] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 2

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 1, 2, 3

[21] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J. Davison. Vmap: Vectorised object mapping for neural field slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 952–961, 2023. 2

[22] Mingrui Li, Zhetao Guo, Tianchen Deng, Yiming Zhou, Yuxiang Ren, and Hongyu Wang. Ddn-slam: Real time dense dynamic neural implicit slam. *IEEE Robotics and Automation Letters*, 10(5):4300–4307, 2025. 2

[23] Lahav Lipson and Jia Deng. Multi-session slam with differentiable wide-baseline pose optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19626–19635, 2024. 2

[24] Lorenzo Liso, Erik Sandström, Vladimir Yugay, Luc Van Gool, and Martin R. Oswald. Loopy-slam: Dense neural slam with loop closures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20363–20373, 2024. 2

[25] Yu Liu, Yilin Wu, and Wenzhao Pan. Dynamic rgb-d slam based on static probability and observation number. *IEEE Transactions on Instrumentation and Measurement*, 70 (8503411):1–11, 2021. 2

[26] 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. 1, 2, 5, 6, 7

[27] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[28] Yuhang Ming, Di Ma, Weichen Dai, Han Yang, Rui Fan, Guofeng Zhang, and Wanzeng Kong. Slc^2-slam: Semantic-guided loop closure using shared latent code for nerf slam. *IEEE Robotics and Automation Letters*, 10(5):4978–4985, 2025. 2

[29] Zhexi Peng, Tianjia Shao, Yong Liu, Jingke Zhou, Yin Yang, Jingdong Wang, and Kun Zhou. Rtg-slam: Real-time 3d reconstruction at scale using gaussian splatting. In *Proceedings of the Annual Conference on Computer graphics and interactive techniques*, pages 1–11, 2024. 2

[30] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 2

[31] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2564–2571, 2011. 2, 4

[32] Erik Sandström, Yue Li, Luc Van Gool, and Martin R. Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18387–18398, 2023. 2

[33] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6209–6218, 2021. 1, 2

[34] 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. 2, 6

[35] Yulun Tian, Hanwen Cao, Sunghwan Kim, and Nikolay Atanasov. Miso: Multiresolution submap optimization for efficient globally consistent neural implicit reconstruction. In *Proceedings of the Robotics: Science and Systems*, 2025. 2

[36] Wenshan Wang, Yaoyu Hu, and Sebastian Scherer. Tartanvo: A generalizable learning-based vo. In *Proceedings of the Conference on Robot Learning*, pages 1761–1772, 2021. 2

[37] 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, 5, 6, 7

[38] Kuan Xu, Yuefan Hao, Shenghai Yuan, Chen Wang, and Lihua Xie. Airslam: An efficient and illumination-robust point-line visual slam system. *IEEE Transactions on Robotics*, 41: 1673–1692, 2025. 2, 3, 4, 5, 6, 7

[39] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19595–19604, 2024. 2

[40] Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-fusion: Dense tracking and mapping with voxel-based neural implicit representation. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, pages 499–507, 2022. 1, 2, 5

[41] Vladimir Yugay, Theo Gevers, and Martin R. Oswald. Magic-slam: Multi-agent gaussian globally consistent slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6741–6750, 2025. 1

[42] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3704–3714, 2023. 2

[43] Jianhao Zheng, Zihan Zhu, Valentin Bieri, Marc Pollefeys, Songyou Peng, and Iro Armeni. Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2025. 1

[44] Wancai Zheng, Xinyi Yu, Jintao Rong, Linlin Ou, Yan Wei, and Libo Zhou. Gsorb-slam: Gaussian splatting slam benefits from orb features and transmittance information. *IEEE Robotics and Automation Letters*, 10(9):9400–9407, 2025. 2

[45] Zhiyu Zhou, Feng Hui, Yilin Wu, and Yu Liu. Six-dof pose estimation with efficient 3-d gaussian splatting representation for visual relocalization. *IEEE/ASME Transactions on Mechatronics*, 30(6):4283–4292, 2025. 1

[46] Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang. Sni-slam: Semantic neural implicit slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21167–21177, 2024. 2

[47] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12776–12786, 2022. 1, 2, 5