IAO-SLAM: Real-time Illumination-Aware Object SLAM for Robust Perception in Low-Light Environments

Pengju Zhen[1] Huilin Yin[1*] Linchuan Zhang[2] Xin Su[3]
[1]The College of Electronic and Information Engineering, Tongji University, Shanghai, China
[2]Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China
[3]The Department of Computer Engineering, CIT, Technical University of Munich, Munich, Germany
*Corresponding author: yinhuilin@tongji.edu.cn

Abstract

Robust perception and navigation remain challenging for visual Simultaneous Localization and Mapping (SLAM) systems in low-light environments, where degraded illumination reduces feature quality and hinders object detection. While geometry-based SLAM under low-light conditions has been explored, most existing solutions either sacrifice real-time performance or neglect the effect of poor illumination on object-level perception and association. In this paper, Illumination-Aware Object SLAM (IAO-SLAM) is proposed, an object-assisted framework explicitly designed to handle low-light scenarios. The system integrates the Zero-DCE++ network and YOLOv12 network to provide real-time low-light image enhancement and object detection. To achieve robust object association, the Adaptive Multi-modal Similarity Fusion (AMSF) strategy combines Wasserstein distance, Intersection over Union (IoU) overlap, and point-feature assistance is introduced, ensuring statistical, geometric, and motion-level consistency. Built upon ORB-SLAM3, our method jointly leverages point features and object landmarks to construct consistent maps in degraded illumination. Comprehensive experiments on the low-light processed TUM RGB-D dataset demonstrate that IAO-SLAM significantly improves localization accuracy and object reconstruction under low-light environments, while also ensuring the real-time performance of the system.

1. Introduction

Visual Simultaneous Localization and Mapping (SLAM) has been extensively studied as a fundamental technology for autonomous perception and navigation in unknown environments. Traditional visual SLAM primarily relies on point or geometric features, which provide accurate localization in well-lit and structured environments. However, such purely geometric pipelines lack semantic understanding of the scene and often struggle to maintain robustness under extreme conditions, limiting their applicability in practical perception and navigation tasks.

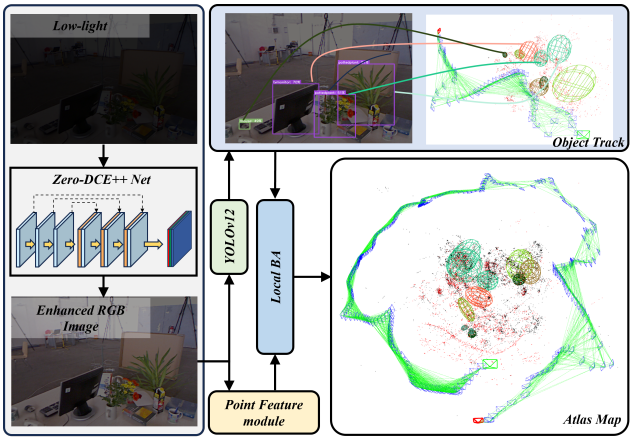


Figure 1. Illustration of low-light image enhancement and object-level mapping. Low-light inputs are first enhanced by Zero-DCE++ network [6], enabling reliable object extraction with YOLOv12 [19]. Enhanced detections are then properly associated with the object map, resulting in atlas maps that jointly contain both feature points and objects.

To overcome the limitations of purely feature-based pipelines, recent researches [11, 17, 18, 21, 25] have extended SLAM with semantic awareness, where objects are introduced as landmarks. Object-aware SLAM typically begins with object detection or instance segmentation to extract semantic instances from images. These objects are then tracked across frames to construct an object-level map, while simultaneously providing additional constraints for camera pose optimization. In this way, object landmarks not only enrich the map with high-level semantics but also

serve as complementary cues to point features, thereby improving robustness for navigation and localization. However, most existing object SLAM systems [21, 25] are not designed for real-time operation. Their frame-wise object detections are often pre-computed and stored offline, which limits their applicability in online robotic systems.

Beyond the real-time limitations of object-based SLAM, low-light conditions degrade feature quality and quantity in geometry-based pipelines, leading to unstable tracking and unreliable pose estimation. At the same time, object detectors may also fail under poor illumination, leading to unreliable associations and incomplete semantic mapping. Several methods [12, 14] focusing on geometry-based SLAM have attempted to mitigate these issues by employing image enhancement networks. While these approaches improve tracking robustness, they remain ineffective under more extreme low-light conditions and rely on models that are not sufficiently lightweight and efficient. Besides, little attention has been paid to the effect of low-light on object extraction and object-level association. This leaves an open problem for developing lightweight illumination-aware object SLAM frameworks that jointly address the degradation of both point features and objects.

To address these challenges, we propose Illumination-Aware Object SLAM (IAO-SLAM), a framework designed to enhance perception and navigation in low-light environments. As shown in Figure 1, the system integrates a lightweight Zero-DCE++ network [6], deployed via the Torch [4] C++ interface, which performs real-time image enhancement while preserving structural details and color fidelity. Besides, an efficient object detection network YOLOv12 [19] is deployed. Furthermore, an Adaptive Multi-modal Similarity Fusion (AMSF) strategy for robust object association is introduced. By combining Wasserstein distance, Intersection over Union (IoU) overlap, and point-feature assistance, AMSF ensures statistical, geometric, and feature-level consistency. Built upon ORB-SLAM3 [2], the system jointly leverages point features and object landmarks, resulting in consistent 3D maps that remain reliable even in challenging low-light scenarios.

The main contributions of this work are summarized as follows:

- A lightweight Zero-DCE++ network [6] and YOLOv12 network [19] is deployed within the SLAM pipeline, enabling real-time low-light enhancement and object detection.
- An Adaptive Multi-modal Similarity Fusion (AMSF) strategy is introduced, which integrates appearance-based Wasserstein distance, geometric IoU overlap, and point-feature assistance to achieve robust object association under degraded illumination.
- Several experiments are conducted to analyze the impact of low-light conditions on system localization and mapping. The results show that our method achieves significant improvements in localization accuracy, object reconstruction, and real-time performance compared to state-of-the-art baselines.

2. Related Works

2.1. Object-aware SLAM Methods

Deep learning-based detection techniques [1, 19, 20] have enabled SLAM to evolve from traditional geometric descriptions toward semantic perception. A lightweight form is object-based SLAM, where objects serve as landmarks, providing compact geometric and semantic information that improves localization and mapping.

Object representations are commonly categorized into learning-based methods, parametric models, and point cloud clusters. By employing multi-view strategies, CubeSLAM [23] reconstructs cuboid boxes and refines object poses. While point cloud clusters methods [8, 24]offer flexibility, they are non-compact and less suitable for high-level tasks.

In contrast, dual quatrics provide a compact parameterization. Building on this property, recent studies [3, 11, 17, 25] have explored integrating quadrics into object-level SLAM, allowing efficient optimization within bundle adjustment frameworks. Nicholson et al. [11] first reconstructed quadrics from 2D bounding boxes, but their method was noise-sensitive. OA-SLAM [25] addresses this with spheres initialization in their own coordinate frames. Based on this, VOOM [21] introduces Wasserstein distance for more robust object association and adopts a coarse-to-fine framework that combines object landmarks with point features to improve odometry and mapping.

Inspired by [25] and [21], we design an Adaptive Multi-modal Similarity Fusion (AMSF) approach for object association. AMSF combines multiple cues, including Wasserstein distance, bounding box IoU, and ellipse shape similarity, into a weighted fusion scheme. This strategy improves robustness to noise and yields more reliable associations in complex environments.

2.2. Illumination-aware SLAM Methods

Visual SLAM methods are highly sensitive to illumination conditions. In low-light environments, insufficient brightness often leads to poor feature extraction and unstable tracking, which directly degrades localization accuracy. Moreover, object detection modules also rely on consistent image quality, and their performance can drop significantly under challenging illumination. Therefore, addressing low-light scenarios is crucial for object-oriented SLAM.

Traditional methods, such as LMVI-SLAM [7], employ adaptive gamma correction and contrast-limited adaptive histogram equalization to improve feature extraction under

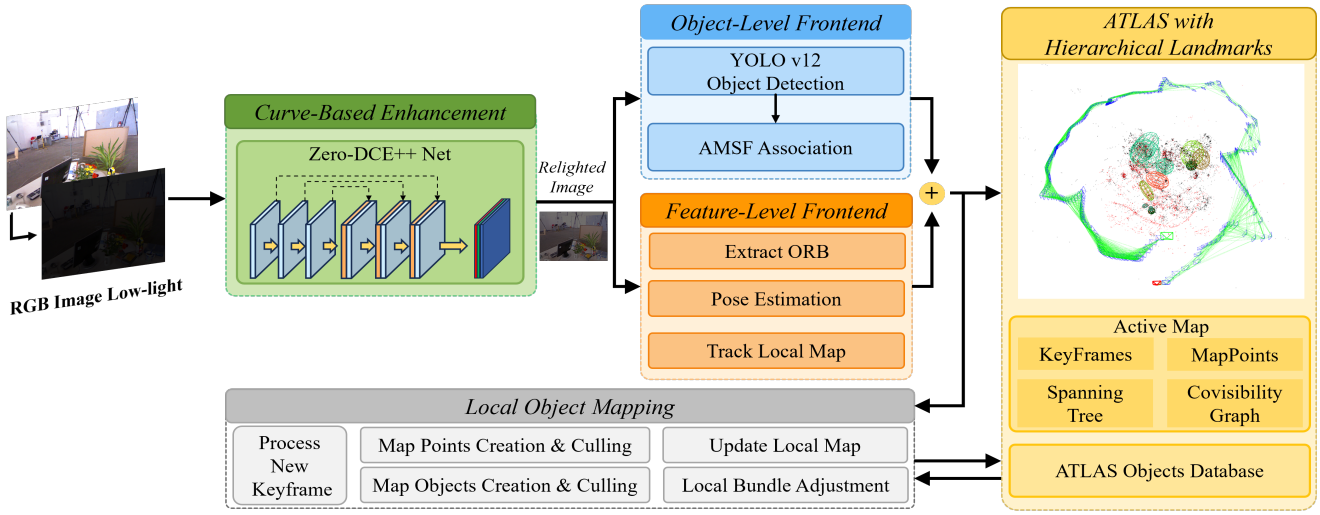


Figure 2. System framework of IAO-SLAM. Low-light images are first enhanced by the curve-based Zero-DCE++ module, followed by feature-level tracking and object-level detection with AMSF-based association. Point features and object landmarks are jointly optimized in local mapping and integrated into the ATLAS for consistent 3D reconstruction.

dim lighting, which remains limited when illumination deteriorates severely. SL-SLAM [22] employs SuperPoint [5] feature extraction together with LightGlue [10] matching, improving robustness in moderately low-light conditions. However, these methods still do not directly enhance the input images. Besides, BVT-SLAM [13] integrates a visible-light camera with a thermal sensor to achieve robust localization in low-light environments, though such sensor fusion inevitably increases hardware complexity and cost.

More recently, deep learning approaches have been explored to enhance low-light images. For instance, [12] applies GAN-based image-to-image translation to mitigate appearance changes, while DarkSLAM [14] integrates EnlightenGAN [9] into the SLAM pipeline for robust tracking in low-light scenes. Twilight-SLAM [15] further benchmarks multiple enhancement models with standard SLAM frameworks. However, most approaches often involve heavy models or do not emphasize deployment efficiency, which limits their applicability in real-time robotic systems. Motivated by these methods, our approach adopts the lightweight Zero-DCE++ [6] network to achieve a better trade-off between efficiency and effectiveness, and further deploys it via ONNX in a C++ environment for practical integration.

3. Methodology

3.1. System Overview

Our system builds upon ORB-SLAM3 [2] and generates a unified map that integrates both point features and object landmarks. As illustrated in Figure 2, the pipeline combines illumination enhancement, feature tracking, object detection and association into a coherent framework. To improve robustness in low-light environments, each incoming frame is pre-processed with the Zero-DCE++ network [6] implemented in Torch [4] C++ interface. Object instances are detected online by YOLOv12 [19], deployed via ONNX Runtime in the C++ environment to ensure real-time performance. In parallel, ORB-SLAM3 extracts and tracks feature points, providing accurate pose prediction. For object association, an Adaptive Multi-modal Similarity Fusion (AMSF) method is employed, where Wasserstein distance, IoU overlap, and point-feature assistance are integrated to ensure statistical, geometric, and feature-level consistency. Point features and objects are jointly integrated into the local Mapping module and the ATLAS system, enabling local map updates and accurate camera poses estimation through bundle adjustment, and ultimately producing a consistent 3D map with both object- and point-level landmarks.

3.2. Curve-Based Low-Light Image Enhancement

To address the problem of low-light images, the Zero-Reference Deep Curve Estimation++ (Zero-DCE++) framework is integrated into the system. Zero-DCE++ formulates enhancement as an image-specific curve estimation without requiring paired or unpaired training data. The method relies on the Light-Enhancement (LE) curve. For a normalized input pixel intensity $I(x) \in [0, 1]$, the quadratic mapping is:

$$LE(I(x); \alpha) = I(x) + \alpha I(x)(1 - I(x)), \qquad (1)$$

where $\alpha \in [-1, 1]$ controls the exposure level. The iterative application of the LE curve provides more flexible adjust-

ment in complex illumination scenarios:

$$LE_n(x) = LE_{n-1}(x) + \alpha_n \cdot LE_{n-1}(x)(1 - LE_{n-1}(x)), \tag{2}$$

where n denotes the iteration index, and it is further extended to pixel-wise parameter maps $A_n(x)$:

$$LE_n(x) = LE_{n-1}(x) + A_n(x) \cdot LE_{n-1}(x)(1 - LE_{n-1}(x)). \tag{3}$$

The parameter maps are predicted by DCE-Net, a lightweight CNN with seven convolutional layers and symmetric skip concatenations. By avoiding down-sampling, the network preserves spatial structures and outputs 24 parameter maps. With approximately 79k trainable parameters and 5.21G FLOPs for an input of size $256 \times 256 \times 3$, DCE-Net can operate in real time on standard hardware.

Training is guided by four non-reference loss functions. The spatial consistency loss preserves local contrast:

$$L_{spa} = \frac{1}{K} \sum_{i=1}^{K} \sum_{j \in \Omega(i)} \left(|Y_i - Y_j| - |I_i - I_j| \right)^2, \tag{4}$$

where Y and I are local averages. The exposure control loss ensures proper brightness:

$$L_{exp} = \frac{1}{M} \sum_{k=1}^{M} |Y_k - E|, \tag{5}$$

where Y_k is the mean intensity of the k-th region and $E = 0.6$. The color constancy loss enforces channel balance:

$$L_{col} = \sum_{\forall (p,q) \in \{(R,G),(R,B),(G,B)\}} (J_p - J_q)^2, \tag{6}$$

where J denotes the average intensity values of a channel. The illumination smoothness loss regularizes parameter maps:

$$L_{tvA} = \frac{1}{N} \sum_{n=1}^{N} \sum_{c \in \{R,G,B\}} \left(|\nabla_x A_n^c| + |\nabla_y A_n^c| \right)^2. \tag{7}$$

The total loss is defined as:

$$L_{total} = L_{spa} + L_{exp} + W_{col} L_{col} + W_{tvA} L_{tvA}. \tag{8}$$

Zero-DCE++ is deployed within our SLAM framework via the Torch [4] C++ interface, and cross-product operations have been adaptively optimized, enabling lightweight and real-time operation. The curve-based formulation robustly enhances low-light images by correcting exposure while preserving details and color fidelity. These properties make it particularly suitable for supporting downstream robotic vision tasks under challenging illumination.

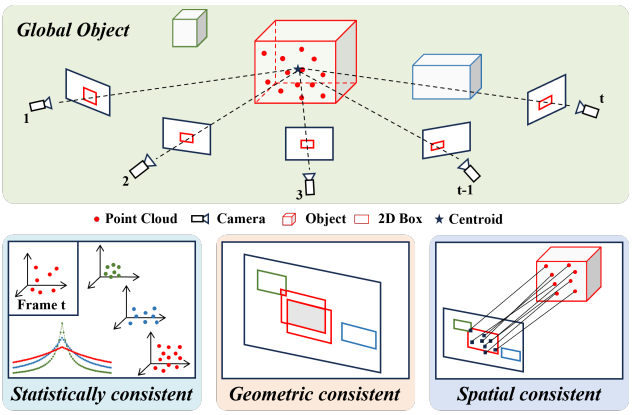


Figure 3. The pipeline of AMSF object association method.

3.3. Adaptive Multi-modal Similarity Fusion (AMSF)

In low-light environments, reliable object tracking becomes particularly challenging, as degraded illumination significantly affects the quality of object detection. To better handle low-light scenarios, an adaptive multi-modal similarity fusion (AMSF) is proposed, as shown in Figure 3, which checks statistical consistency, geometric consistency, and Spatial consistency.

Each object is modeled as a 3D ellipsoid, represented by a symmetric dual matrix $Q^* \in \mathbb{R}^{4\times4}$ as follows:

$$Q^* = \begin{bmatrix} A & b \\ b^\top & c \end{bmatrix}, \tag{9}$$

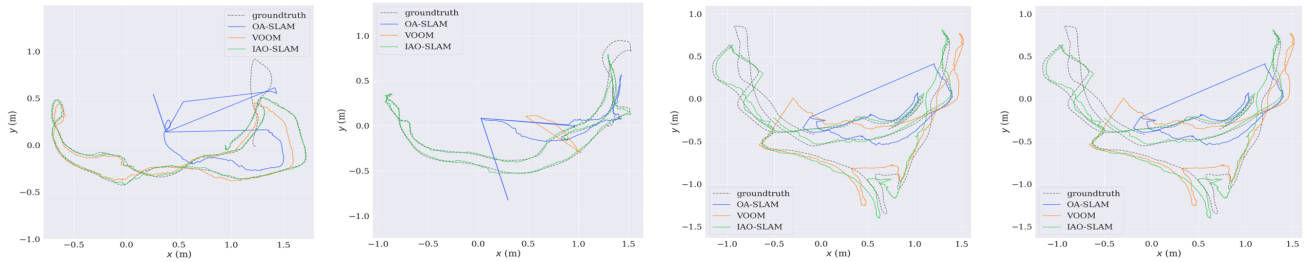
where $A \in \mathbb{R}^{3\times3}$ encodes the orientation and scale, $b \in \mathbb{R}^{3\times1}$ encodes the translation, and c is a scalar term. Given the camera projection matrix $P_f = K_f \left[R_f \mid t_f \right]$, where K_f is the intrinsic calibration and (R_f, t_f) are extrinsics, the ellipsoid is projected onto the image plane as a dual conic C_f^*:

$$C_f^* = P_f Q^* P_f^T \in \mathbb{R}^{3\times3}. \tag{10}$$

This projection corresponds to an ellipse, which is compared against the detected object appearance in the image. For robustness under low-light conditions, the observation ellipse is interpreted as a Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. To robustly compare an observation ellipse E_f^{obs} with a projection E_f^{est}, we employ the 2-Wasserstein distance between two Gaussian distributions:

$$W_2^2(\mathcal{N}_1, \mathcal{N}_2) = \|\mu_1 - \mu_2\|_2^2 + \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_F^2, \tag{11}$$

where μ is the ellipse center, Σ encodes its axes and orientation, and $\|\cdot\|_F$ denotes the Frobenius norm. Since the raw Wasserstein distance is sensitive to object scale, we further



(a) fr1_desk (b) fr1_desk2 (c) fr1_room (d) fr2_desk

Figure 4. The comparison of estimated trajectories of IAO-SLAM, OA-SLAM, VOOM and Ground Truth on the low-light processed TUM RGB-D dataset.

adopt a normalized Wasserstein score:

$$S(E_f^{\text{obs}}, E_f^{\text{est}}) = \exp\left(-\frac{\sqrt{W_2^2(\mathcal{N}_{\text{obs}}, \mathcal{N}_{\text{est}})}}{C}\right), \quad (12)$$

where C is a normalization constant, enabling stable data associations. This normalized distance guarantees that associations remain statistically consistent.

To further constrain associations, we incorporate the Intersection-over-Union (IoU) between the detected bounding box B_f^{det} and the projection of the ellipsoid $\text{proj}(Q^*)$:

$$\text{IoU}(B_f^{\text{det}}, \text{proj}(Q^*)) = \frac{|B_f^{\text{det}} \cap \text{proj}(Q^*)|}{|B_f^{\text{det}} \cup \text{proj}(Q^*)|}, \quad (13)$$

where the numerator and denominator are the intersection area and the union area, respectively. This ensures geometric overlap consistency, especially when illumination reduces segmentation quality.

Finally, associations are stabilized by checking point-feature consistency. A detection D_f and an ellipsoid Q^* are matched if sufficient correspondences satisfy $x_i \in D_f, \quad X_i \in Q^*$, where x_i is an image keypoint inside the detection box and X_i its 3D landmark inside the ellipsoid. The consistency score is defined as

$$\delta(D_f, Q^*) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big(x_i \in D_f \wedge X_i \in Q^*\big), \quad (14)$$

where N is the number of matches and $\mathbf{1}(\cdot)$ is an indicator function. This measure validates feature-level accuracy to reinforce temporal and spatial coherence.

4. EXPERIMENT

In this section, we present the experimental evaluation of our system. The localization accuracy against state-of-the-art methods is evaluated on the 6 sequences of TUM RGB-D dataset[16]. The perception front-end employs YOLOv12, which is deployed in the C++ environment using ONNX Runtime to achieve real-time object detection. To handle low-light scenarios, the Zero-DCE++ enhancement model is integrated through a Torch C++ interface implementation. All implementations are developed in C++ on Ubuntu 20.04, and the experiments are conducted on a laptop equipped with an Intel Core i7-12700H @ 2.70GHz CPU, 16GB RAM, and an NVIDIA GeForce RTX 3060 GPU. The root mean square error (RMSE) is used to analyze the localization accuracy and stability of our method.

To evaluate robustness under challenging illumination conditions, low-light image sequences are synthetically generated by adaptively attenuating the brightness of the original dataset. A multi-peak Gaussian decay model is designed to control the frame-wise brightness scaling factor α_i, which is normalized to $[0, 1]$. That ensures smooth yet non-uniform illumination degradation across the sequence.

To further approximate realistic sensor behavior under poor lighting, stochastic noise is added after Gaussian decay. The noise is sampled by sparse random samples drawn from a uniform distribution and then interpolated linearly to form a continuous curve. In this way, the simulated brightness exhibits fluctuations similar to unstable illumination conditions in real-world environments.

4.1. Experiments on Low-light Conditions

Table 2 summarizes the localization accuracy of ORB-SLAM3, OA-SLAM, VOOM and the proposed IAO-SLAM on the TUM RGB-D dataset under both normal illumination and low-light conditions. For each method, the value before "/" indicates the localization error under normal illumination, the value after "/" corresponds to the error under low-light conditions. To enable comparisons on object-related SLAM, we extend the original offline-detection versions of OA-SLAM and VOOM to online detection using YOLOv12.

Across all sequences, introducing low-light severely degrades performance for conventional approaches. For example, the error of ORB-SLAM3 on fr2_desk increases from 0.017 to 0.648, while OA-SLAM fails in several cases (e.g., fr1_room, fr2_desk) with errors exceeding 1.5. Similarly, VOOM also suffers from unstable results, where the

Table 1. RMSE APE comparison on TUM-RGB-D dataset (unit: m). For each table, the localization error of the original method is shown before "/", and the error after conducting low-light is shown after "/". ↑ depicts that the localization error increases after introducing complex low-light. ↓ indicates that the error decreases. "-" represents tracking failure. Numbers in bold represent the minimum localization error.

Methods		fr1_desk	fr1_desk2	fr1_room	fr1_xyz	fr2_desk	fr2_xyz
ORB-SLAM3	Mean	0.017/0.034↑	0.218/0.275↑	0.069/0.541↑	**0.010**/0.011↑	**0.011**/0.648↑	**0.003**/0.006↑
	Std	0.010/0.025↑	**0.010**/0.016↑	0.030/0.187↑	0.006/0.006	0.005/0.378↑	0.002/0.005↑
OA-SLAM[1]	Mean	0.693/0.768↑	0.507/0.762↑	0.999/0.867↓	0.108/0.156↑	0.227/1.579↑	0.041/0.151↑
	Std	0.374/0.303↓	0.334/0.457↑	0.356/0.475↑	0.082/0.071↓	0.198/0.506↑	0.040/0.083↑
VOOM[2]	Mean	0.019/0.201↑	0.027/-↑	0.136/0.617↑	0.016/0.140↑	0.023/0.241↑	0.008/0.073↑
	Std	0.012/0.143↑	0.013/-↑	0.050/0.380↑	0.008/0.063↑	0.014/0.191↑	0.004/0.063
IAO-SLAM	Mean	0.016/**0.015**↓	0.023/**0.022**↓	**0.061**/0.069↑	**0.010/0.010**	**0.011**/0.018↑	**0.003**/0.004↑
	Std	0.009/**0.008**↓	0.012/0.011↓	0.029/**0.028**↓	**0.005**/0.006↑	**0.004**/0.009↑	**0.001**/0.002↑

[1.] The offline object detection in 1 and 2 is replaced with the same YOLOv12-based online processing as in the proposed method.

Table 2. Ablation study on the TUM-RGB-D dataset (unit: m). RMSE and standard deviation (Std) of localization errors are reported.

Sequence	Metric	IAO-SLAM w/o Object	IAO-SLAM w/o Zero-DCE++	IAO-SLAM
fr1_desk	RMSE	0.016	0.022	0.015
	Std	0.009	0.010	0.008
fr1_desk2	RMSE	0.038	0.764	0.026
	Std	0.016	0.325	0.011
fr1_room	RMSE	0.091	0.348	0.068
	Std	0.047	0.217	0.028
fr2_desk	RMSE	0.018	0.0238	0.011
	Std	0.010	0.169	0.009

error on fr1_room rises to 0.136. These results highlight the strong sensitivity of geometry- and object-based SLAM pipelines to illumination degradation.

In contrast, the proposed IAO-SLAM demonstrates notably higher robustness. Even without image enhancement, the error increase under low-light conditions remains moderate across all sequences. As evidenced by the trajectories in Figure 4, IAO-SLAM maintains close alignment with the ground truth, while OA-SLAM and VOOM exhibit significant drift or divergence. As also reported in Table 2, IAO-SLAM consistently achieves the lowest errors, effectively reducing the degradation caused by illumination loss. These improvements demonstrate that the proposed design not only stabilizes object detection but also enhances overall trajectory accuracy. At the same time, the results indicate that the proposed method does not compromise localization accuracy under normal illumination conditions.

4.2. Ablation Study

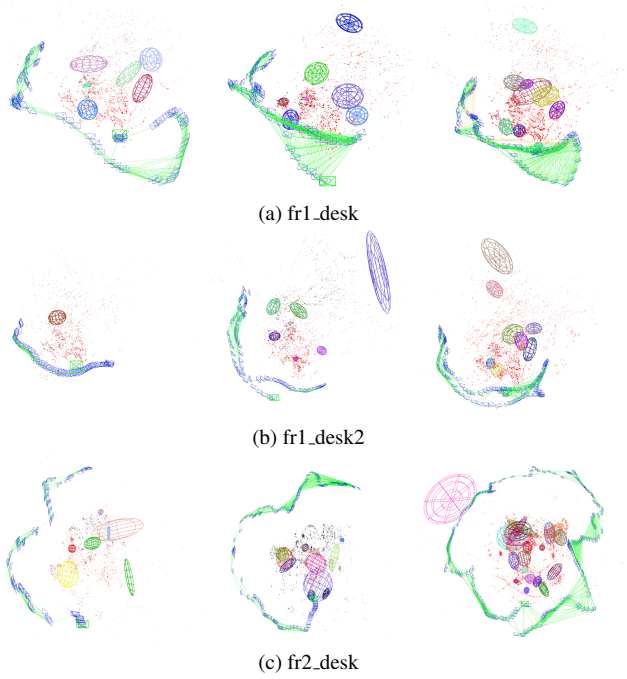
Ablation results on the TUM RGB-D dataset are summarized in Table 2, where RMSE and standard deviation (Std) are reported for different configurations. Specifically, we compare the full IAO-SLAM system with two degraded variants, i.e., removing the object-related module (IAO-SLAM w/o Object) and removing the low-light enhancement module (IAO-SLAM w/o Zero-DCE++). As shown in the table, introducing object-level constraints consistently improves localization accuracy. The full system achieves lower RMSE and error variance than IAO-SLAM w/o Object across all sequences, with more noticeable improvements in challenging sequences such as fr1_desk2 and fr1_room, indicating that stable object association provides more reliable structural information.

In addition, the low-light enhancement module plays a critical role in maintaining robustness. Without Zero-DCE++, the localization error increases significantly in several sequences, especially fr1_desk2 and fr1_room, where both RMSE and Std show substantial degradation. In contrast, the full IAO-SLAM achieves the best overall performance, demonstrating that low-light enhancement effectively improves feature quality and stabilizes object detection, thereby enhancing the overall localization accuracy under challenging lighting conditions.

4.3. Object-Related Improvement

In addition to localization accuracy, we further investigate the impact of low-light conditions on object-level mapping. As illustrated in Figure 5, OA-SLAM and VOOM often suffer from incomplete or distorted object reconstructions when the image quality is degraded. In particular, OA-SLAM fails to maintain stable associations, resulting in fragmented ellipsoids, while VOOM exhibits unstable detections and sparse object landmarks.

In contrast, the proposed IAO-SLAM generates significantly more consistent and complete object maps under the same low-light conditions. The objects are not only de-



(a) fr1_desk

(b) fr1_desk2

(c) fr2_desk

Figure 5. Map comparison on three TUM-RGBD dataset sequences under low-light conditions. For each sequence, the figures from left to right present the mapping results of OA-SLAM, VOOM, and IAO-SLAM, respectively.

tected more reliably but also represented with higher geometric consistency, which directly contributes to improved map quality. This improvement stems from the adaptive multi-feature fusion and the proposed object association strategy, which together ensure robustness against degraded visual cues.

The quantitative results in Table 2 further support these findings. While OA-SLAM shows severe error growth and even tracking failures in low-light sequences, IAO-SLAM achieves the lowest errors across all sequences, confirming that reliable object-level perception complements trajectory estimation and enhances robustness in challenging environments.

4.4. Real-time Performance

To evaluate the efficiency of the proposed system, we analyze the run-time performance of the Zero-DCE++ image enhancement module, YOLOv12 object detection module and AMSF tracking module. Table 3 presents the average execution time of these three modules, evaluated on four representative sequences from the TUM RGB-D dataset. The lightweight Zero-DCE++ network, deployed via the Torch C++ interface, achieves an average processing time of within 13–14 ms per frame, confirming that the network ensures efficient processing without noticeable delay. Similarly, the YOLOv12-based object detector requires about

Table 3. Real-time performance. Average Execution Time of different modules run on four sequences of TUM RGB-D Dataset.

Module	Running Time (ms)			
	fr1_desk	fr1_desk2	fr1_room	fr2_desk
Zero-DCE++ Image Enhancement	13.4	13.6	13.8	13.5
Object Detection	9.5	9.5	9.6	9.4
AMSF Tracking	1.2	1.1	2.4	3.4

9–10 ms per frame, showing that reliable online detection can be achieved within the real-time budget. AMSF tracking module exhibits slightly varying runtimes depending on the number of objects present in the scene, ranging from 1-3ms, yet it still contributes only a negligible overhead to the overall pipeline.

These results indicate that the integration of the Zero-DCE++ network, the YOLOv12 network and the AMSF tracking module only introduces less than 30 ms of latency per frame, which satisfies real-time requirements and supports practical applications.

5. CONCLUSIONS

In this work, IAO-SLAM is proposed, an illumination-aware object SLAM framework for perception and navigation in low-light environments. The lightweight Zero-DCE++ network and YOLOv12 network can achieve real-time image enhancement and object detection, respectively. The AMSF strategy is introduced to ensure statistical, geometric, and motion-level consistency. It can jointly exploit point and object landmarks to construct consistent 3D maps under degraded illumination. Experiments on the low-light processed TUM RGB-D dataset demonstrated that IAO-SLAM improves localization accuracy and object-level map construction, while maintaining real-time operation. These results highlight the importance of illumination-aware designs in SLAM. Future work will extend IAO-SLAM to real-world outdoor scenarios with dynamic lighting.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 62433014 and No.62133011. This work was also supported by the State Key Laboratory of Autonomous Intelligent Unmanned Systems and Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education of China. We are also grateful for the efforts from our colleagues in Sino German Center of Intelligent Systems. This work also has been partially supported by the German Federal Ministry of Research, Technology and Space (BMFTR) under the Robotics Institute Germany (RIG).

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 2

[2] 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, 3

[3] ZhenZhong Cao, Yunzhou Zhang, Rui Tian, Rong Ma, Xinggang Hu, Sonya Coleman, and Dermot Kerr. Object-aware slam based on efficient quadric initialization and joint data association. *IEEE Robotics and Automation Letters*, 7(4): 9802–9809, 2022. 2

[4] Ronan Collobert, Samy Bengio, and Johnny Mariéthoz. Torch: a modular machine learning software library. 2002. 2, 3, 4

[5] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 3

[6] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1777–1786, 2020. 1, 2, 3

[7] Luoying Hao, Hongjian Li, Qieshi Zhang, Xiping Hu, and Jun Cheng. Lmvi-slam: Robust low-light monocular visual-inertial simultaneous localization and mapping. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 272–277, 2019. 2

[8] Jiahui Huang, Sheng Yang, Tai-Jiang Mu, and Shi-Min Hu. Clustervo: Clustering moving instances and estimating visual odometry for self and surroundings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2168–2177, 2020. 2

[9] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021. 3

[10] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17627–17638, 2023. 3

[11] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 4(1):1–8, 2018. 1, 2

[12] Horia Porav, Will Maddern, and Paul Newman. Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1011–1018. IEEE, 2018. 2, 3

[13] Liang Qin, Chang Wu, Xiaotong Kong, Yuan You, and Zhiqi Zhao. Bvt-slam: a binocular visible-thermal sensors slam system in low-light environments. *IEEE Sensors Journal*, 24 (7):11599–11609, 2023. 3

[14] Alena Savinykh, Mikhail Kurenkov, Evgeny Kruzhkov, Evgeny Yudin, Andrei Potapov, Pavel Karpyshev, and Dzmitry Tsetserukou. Darkslam: Gan-assisted visual slam for reliable operation in low-light conditions. In *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pages 1–6. IEEE, 2022. 2, 3

[15] Surya Pratap Singh, Billy Mazotti, Dhyey Manish Rajani, Sarvesh Mayilvahanan, Guoyuan Li, and Maani Ghaffari. Twilight slam: Navigating low-light environments. *arXiv preprint arXiv:2304.11310*, 2023. 3

[16] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 573–580, 2012. 5

[17] Rui Tian, Yunzhou Zhang, Yonghui Feng, Linghao Yang, Zhenzhong Cao, Sonya Coleman, and Dermot Kerr. Accurate and robust object slam with 3d quadric landmark reconstruction in outdoors. *IEEE Robotics and Automation Letters*, 7(2):1534–1541, 2021. 1, 2

[18] Rui Tian, Yunzhou Zhang, Linghao Yang, Jinpeng Zhang, Sonya Coleman, and Dermot Kerr. Dynaquadric: Dynamic quadric slam for quadric initialization, mapping, and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):17234–17246, 2024. 1

[19] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 1, 2, 3

[20] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024. 2

[21] Yutong Wang, Chaoyang Jiang, and Xieyuanli Chen. Voom: Robust visual object odometry and mapping using hierarchical landmarks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10298–10304, 2024. 1, 2

[22] Z Xiao and S Li. Sl-slam: A robust visual-inertial slam based deep feature extraction and matching. *arXiv preprint arXiv:2405.03413*, 2024. 3

[23] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35(4):925–938, 2019. 2

[24] Jun Zhang, Mina Henein, Robert Mahony, and Viorela Ila. Vdo-slam: A visual dynamic object-aware slam system. *arXiv preprint arXiv:2005.11052*, 2020. 2

[25] Matthieu Zins, Gilles Simon, and Marie-Odile Berger. OA-SLAM: Leveraging objects for camera relocalization in visual slam. In *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 720–728, 2022. 1, 2