

LUNA: Low-Light Robust Panoptic Lifting for Adverse Robotic 3D Scene Perception

Ahalya Ravendran¹, Xun Li¹, Léo Lebrat², Rodrigo Santa Cruz², Hu Zhang¹, Lars Petersson¹ and Dadong Wang¹

Abstract—Robotic perception in low-light environments remains challenging due to severe noise and motion blur that degrade RGB imagery and hinder reliable scene understanding. Although recent 3D panoptic reconstruction methods unify semantic and instance segmentation, they largely assume high-quality inputs - an assumption that fails under real-world conditions. We present LUNA, a geometry-aware panoptic lifting framework that integrates depth-guided consistency and adaptive multitask optimization for robust reconstruction under degraded visual inputs. We further introduce a systematically degraded Replica dataset with progressive noise and blur to evaluate robustness. Experiments show that LUNA consistently outperforms restoration-based baselines in panoptic segmentation and volumetric reconstruction across all degradation levels.

I. INTRODUCTION

Robots deployed in real-world environments must be capable of perceiving and understanding complex scenes under diverse visual conditions [1]. Tasks such as autonomous navigation, exploration, and search-and-rescue require comprehensive 3D scene understanding that supports both geometric and semantic reasoning. Recent advances have extended 3D reconstruction methods such as neural radiance field (NeRF) [2] and Gaussian splatting [3] towards panoptic reconstruction frameworks [4]–[6], which jointly capture geometry and semantics for more holistic scene representations.

However, images captured in real-world robotic scenarios often suffer from low-light, resulting in high noise and motion blur, particularly in environments such as underground mines, disaster zones, or during night operations [7]. These degradations significantly impair robust 3D scene understanding. While low-light reconstruction has received increasing attention [8]–[10], panoptic reconstruction remains unexplored in such settings. Existing methods rely heavily on RGB cues, which become unreliable under severe image degradations, leading to inaccurate semantic predictions and geometrically inconsistent reconstructions [4], [6].

In this work, we propose LUNA, a geometry-aware panoptic lifting framework designed to operate under low-light, noisy, and blurred imaging conditions. Built upon panoptic lifting [4], LUNA introduces geometry-aware supervision optimized using fast adaptive multitask optimization (FAMO) [11]. We leverage depth supervision to enhance spatial consistency and

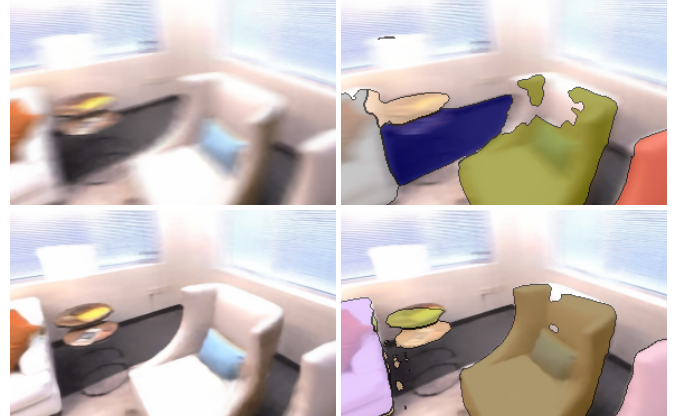


Fig. 1. Example scene from the Replica dataset demonstrating motion blur degradation. Top-left: Reconstructed novel view under blur using panoptic lifting. Bottom-left: Reconstruction from the same input using our method (LUNA), recovering finer structural and texture details. Top-right: Instance segmentation by panoptic lifting misses key objects (e.g., chair, shelf). Bottom-right: LUNA restores missing instances and improves segmentation quality, showing clearer object boundaries and improved completeness.

semantic fidelity, while FAMO automatically balances these objectives during training. We evaluate LUNA across multiple levels of simulated noise and blur on the Replica dataset [12], with ablation studies analyzing each loss component. Fig. 1 illustrates the improved reconstruction and segmentation using LUNA under blur degradation. Based on these findings, we also provide recommendations for adapting LUNA to robotic settings. Upon acceptance, the code and dataset will be made publicly available.

Our main contributions are as follows:

- We present **LUNA**, a unified framework for 3D scene understanding under adverse visual conditions, incorporating geometry-aware supervision that leverages depth information to enhance spatial consistency and semantic fidelity.
- We integrate **fast adaptive multitask optimization (FAMO)** into the training process to automatically balance geometric and semantic objectives, and propose a systematic evaluation protocol using a degraded Replica dataset with controlled levels of noise and blur to assess robustness under challenging conditions.
- We demonstrate that LUNA consistently outperforms panoptic lifting and restoration-augmented baselines in both segmentation and reconstruction tasks across all

¹The authors are with the Imaging and Computer Vision Research Group, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia ahalya.ravendran@data61.csiro.au

²Léo Lebrat and Rodrigo Santa Cruz are with the School of Electrical Engineering and Robotics, Queensland University of Technology (QUT), Australia.

degradation levels, and we provide actionable recommendations for deploying LUNA in robotic settings.

II. RELATED WORK

Low-Light Neural Rendering. Neural rendering under visual challenges has become an active area of research, with several recent works proposing strategies to improve scene representations such as radiance fields and point-based splatting under extreme lighting conditions. RawNeRF [9], LLeNeRF [8], GloNeRF [10], HDR-GS [13], and Gaussian-DK [14] all aim to reconstruct high-fidelity scenes from dark or HDR environments. These methods commonly rely on RAW sensor data or high-dynamic-range priors to stabilize appearance rendering. While these methods achieve state-of-the-art performance in photorealistic rendering under low-light, they are appearance-centric and lack integration with downstream perception tasks such as semantic or panoptic segmentation. As a result, they fall short for robotics applications, where structured scene understanding is critical beyond producing visually plausible output.

Image Restoration for Robotics. Traditional low-light image enhancement has explored both single-image and burst-mode approaches [15], [16] to mitigate noise and blur. For instance, SDAP [17] presents a self-supervised denoising framework utilizing pseudo-labels from sampling differences, while LaKDNet [18] introduces a lightweight deblurring network optimized for edge devices. Although effective in restoring image fidelity, these methods focus primarily on visual quality rather than geometric consistency. As a result, applying them directly in vision pipelines for 3D reconstruction or segmentation can introduce geometric inconsistencies [7], [19], particularly when used as preprocessing for structure-from-motion or neural rendering pipelines. For robotic vision, where geometric accuracy is essential, these methods offer limited utility.

Panoptic Reconstruction in Robotics. Panoptic segmentation has emerged as a critical capability in 3D scene understanding, enabling unified semantic and instance-level segmentation. Systems like Panoptic Fusion [20] and Panoptic Multi-TSDF [21] integrate 2D segmentation into 3D volumetric maps for robotic navigation. However, these systems rely on handcrafted 2D-3D fusion logic and are not optimized for end-to-end training. Panoptic lifting [4] advances this paradigm by unifying neural radiance fields with volumetric panoptic segmentation through a differentiable rendering pipeline. Despite its innovation, this assumes high-quality RGB inputs and is vulnerable to appearance degradations common in real-world robotic scenarios.

LUNA bridges this gap by enhancing the robustness of panoptic lifting with geometry-aware supervision optimized through FAMO [11]. Unlike prior works that either focus on image restoration or volumetric segmentation in isolation, LUNA integrates depth supervision directly into the rendering pipeline and learns to balance loss terms adaptively across multiple tasks. This design enables LUNA to maintain geometric fidelity and semantic consistency under adverse conditions,

TABLE I

COMPARISON OF CAPABILITIES ACROSS RECENT METHODS. ✓ INDICATES SUPPORTED FUNCTIONALITY. LUNA SUPPORTS LOW-LIGHT PERCEPTION, PANOPTIC SEGMENTATION, AND 3D RECONSTRUCTION WITHIN A UNIFIED FRAMEWORK.

Method	3D Reconstruction	Segmentation	Low-Light
Nice-SLAM	✓	✗	✗
Semantic NeRF	✗	✓	✗
DarkNeRF	✗	✗	✓
Panoptic Lifting	✓	✓	✗
LUNA (Ours)	✓	✓	✓

such as noise and motion blur. To the best of our knowledge, LUNA is the first framework to explicitly address panoptic volumetric perception under low-light conditions (see Table I), making it directly applicable to field robotics operating in visually degraded environments.

III. METHOD

We introduce LUNA, a robust extension of panoptic lifting [4] that addresses the failure modes of purely appearance-driven volumetric segmentation under severe noise and motion blur. Like panoptic lifting, LUNA employs a backbone leveraging Mask2Former [22] and TensorRF [23] to jointly predict 3D semantic and instance labels by rendering volumetric features from RGB images and projecting panoptic masks back into the 3D volume. However, under degraded visual conditions, RGB cues alone become unreliable. To address this, LUNA integrates two complementary components: (1) Geometry-aware depth supervision, optimized using FAMO [11], to enforce detailed 3D geometry, improve photometric fidelity, and ensure both semantic and instance consistency; and (2) an optional image-restoration front-end for analyzing the effect of denoising or deblurring on reconstruction quality. The quantitative impact of these preprocessing variants is reported separately in IV-C. An overview of the complete LUNA pipeline is shown in Fig. 2. Below, we first summarize the volumetric rendering formulation from [4] for completeness, then describe our extensions.

A. Volumetric Rendering and Feature Integration

Following panoptic lifting [4], we represent the scene as a panoptic radiance field $\Phi(\mathbf{x}, \mathbf{d})$ that maps each 3D point $\mathbf{x}$ and view direction $\mathbf{d}$ to a density σ , color $\mathbf{c}$, semantic class distribution κ , and instance (surrogate identifier) distribution π . This allows the model to jointly encode appearance, semantics, and instances in a unified volumetric representation.

Given the density field σ from Φ , we can render any vector field f (such as color or class probabilities) along a camera ray r as

$$\hat{\mathbf{f}}_r = \sum_{z \in Z_r} \alpha_z \mathbf{f}_z, \quad (1)$$

where α_z denotes the opacity weight at depth z .

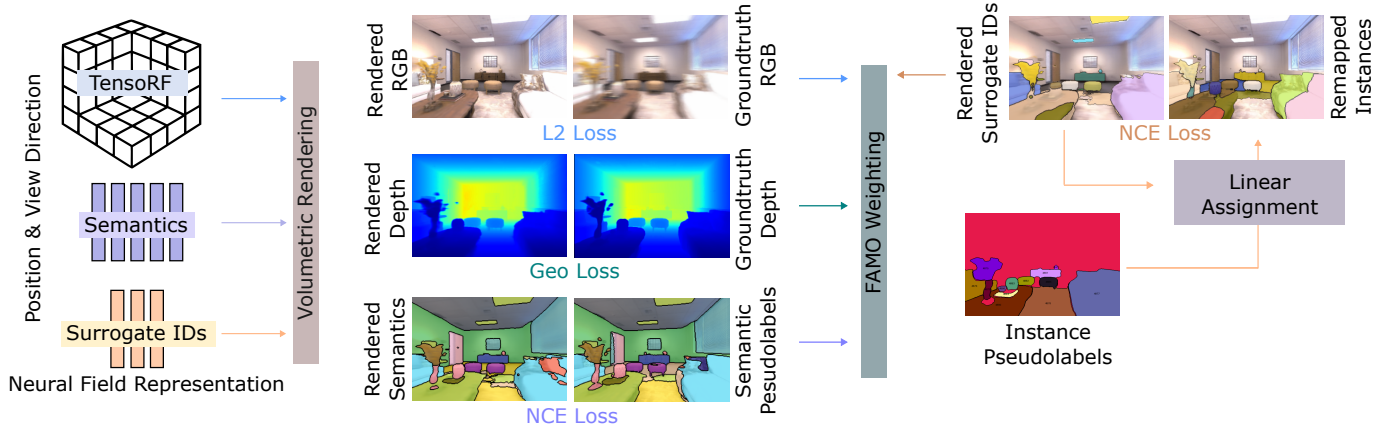


Fig. 2. Overview of the LUNA framework. Multi-view RGB images degraded by low-light, noise, or blur are processed by a volumetric renderer built on TensorRF with a Mask2Former head for panoptic prediction. Geometry-aware depth supervision forms a composite geometric loss L_{geo} that combines absolute depth error and smoothness. Photometric, geometric, semantic, and instance objectives are treated as separate tasks within FAMO, which dynamically weights each loss via a softmax over learnable meta-weights. The network parameters and FAMO weights are jointly trained end-to-end under distributed data parallelism (DDP).

From these rendered features, we obtain the RGB color $\hat{I}_r$, semantic logits $\hat{\mathbf{p}}_r$, and instance masks $\hat{\mathbf{m}}_r$. The rendered depth is similarly defined as

$$\hat{d}_r = \sum_{z \in \mathbb{Z}_r} \alpha_z z. \quad (2)$$

This unified rendering formulation provides 3D-consistent panoptic predictions across views and serves as the foundation for our geometry-aware and multitask extensions described in the following sections.

B. Geometry-Aware Depth Supervision

In panoptic lifting [4], the network is trained using photometric, semantic, instance, and segment-consistency losses. The photometric loss L_{RGB} is an ℓ_2 reconstruction term between the rendered color $\mathbf{c}_r$ and ground-truth color $\hat{\mathbf{c}}_r$,

$$L_{\text{RGB}} = \frac{1}{|R|} \sum_{r \in R} \|\mathbf{c}_r - \hat{\mathbf{c}}_r\|_2^2, \quad (3)$$

while the semantic and instance losses supervise the volumetrically rendered class probabilities κ_r and instance identifiers π_r against 2D machine-generated labels via cross-entropy and log-loss objectives, respectively. A lightweight segment-consistency loss further promotes spatially coherent segmentations.

To enhance 3D reconstruction fidelity under degraded inputs, we introduce an additional *geometry-aware depth supervision* loss $\mathcal{L}_{\text{geo}}$ that complements these appearance-driven terms:

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{depth}} \mathcal{L}_{\text{abs}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}, \quad (4)$$

where $\mathcal{L}_{\text{abs}}$ enforces global depth accuracy by measuring the absolute depth error using an $\mathcal{L}_1$ loss and $\mathcal{L}_{\text{grad}}$ promotes local smoothness by enforcing gradient consistency between neighboring depth values to promote edge-preserving reconstructions,

$$\mathcal{L}_{\text{abs}} = \frac{1}{|R|} \sum_{r \in R} |d_r - \hat{d}_r|, \quad (5)$$

$$\mathcal{L}_{\text{grad}} = \frac{1}{|R|} \sum_{r \in R} |\nabla d_r - \nabla \hat{d}_r|. \quad (6)$$

Here d_r and $\hat{d}_r$ denote the ground-truth and rendered depths along ray r , and ∇ is the spatial gradient operator. This formulation jointly encourages metric depth precision and edge preservation, improving robustness to noise, blur, and low illumination. Note that the ground-truth depths may originate from any modality such as LiDAR or stereo and are converted to depth maps compatible with the volumetric renderer.

C. Multi-Loss Optimization with FAMO

In panoptic lifting, all objectives are combined through manually chosen scalar weights,

$$L_{\text{tot}} = \lambda_{\text{RGB}} L_{\text{RGB}} + \lambda_{\text{sem}} L_{\text{sem}} + \lambda_{\text{ins}} L_{\text{ins}} + \lambda_{\text{con}} L_{\text{con}}. \quad (7)$$

Such static weighting can bias optimization toward dominant tasks. In contrast, LUNA employs FAMO [11] to dynamically balance all objectives including our new geometric loss. At each iteration, we compute the task-specific losses

$$\ell = [L_{\text{RGB}} \quad L_{\text{sem}} \quad L_{\text{ins}} \quad L_{\text{con}} \quad L_{\text{geo}}]^\top, \quad (8)$$

where L_{RGB} is the photometric reconstruction loss between the input and rendered images; L_{sem} is the semantic segmentation loss, computed as pixel-wise cross-entropy over rendered class logits; L_{ins} is the instance segmentation loss based on binary mask IoU; L_{con} is the segment consistency loss that groups rays according to their 2D instance or stuff class; and L_{geo} is the geometric loss defined in Eq. 4.

Each task is associated with a learnable meta-weight w_i , forming $\mathbf{w} = [w_{\text{RGB}}, w_{\text{sem}}, w_{\text{ins}}, w_{\text{con}}, w_{\text{geo}}]^\top$. Normalized task weights are obtained via

$$\alpha_i = \frac{\exp(w_i)}{\sum_{j=1}^5 \exp(w_j)}, \quad (9)$$

$$L_{\text{FAMO}} = \sum_{i=1}^5 \alpha_i \ell_i. \quad (10)$$

During back-propagation, FAMO updates w through a meta-gradient step that maximizes cosine similarity among task gradients, preventing any single objective from dominating the optimization direction. This unified, geometry-augmented formulation yields stable convergence under severe visual degradation and removes the need for manual loss scheduling.

D. Implementation Details

LUNA is implemented in PyTorch following the panoptic lifting [4] scheduler, with learning rates of 5×10^{-4} for MLPs, 1×10^{-2} for TensoRF components, and a FAMO meta-learning rate of 1×10^{-3} . We retain the volumetric rendering backbone and Mask2Former segmentation head, and train the model under *Distributed Data Parallelism (DDP)* on four NVIDIA H100 GPUs for 576,000 iterations, achieving convergence in approximately four hours.

IV. RESULTS

A. Synthetic Dataset Generation

We construct a systematically degraded version of the Replica dataset [12] using six indoor scenes to simulate realistic visual degradations encountered in low-light robotic scenarios. We apply two types of degradations to the 1200×680 RGB images: additive Gaussian noise and motion blur. Specifically, Gaussian noise with standard deviations of $\sigma = 40, 80, 120$ corresponds to the *Noise-Easy*, *Noise-Moderate*, and *Noise-Strong* conditions, respectively. Motion blur is synthesized by convolving each image with a uniform blur kernel of size $k = 10, 20, 30$ pixels, simulating *Blur-Easy*, *Blur-Moderate*, and *Blur-Strong* levels. Fig. 3 illustrates representative examples of the degraded RGB images alongside their original clean counterpart and corresponding ground-truth depth map.

B. Evaluation Metrics

We evaluate our method in terms of both reconstruction quality and panoptic segmentation accuracy. For reconstruction quality, we report peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) on synthesized novel views. For panoptic segmentation, we follow [4], [24] and adopt mean intersection over union (mIoU) and panoptic quality (PQ) as primary metrics. PQ is further decomposed into Segmentation Quality (SQ) and Recognition Quality (RQ), where SQ measures the average IoU of matched segments and RQ behaves as an F1-score-like measure of detection accuracy that balances true positives, false positives, and false negatives. This decomposition provides insight into whether performance degradation under adverse conditions is primarily due to errors in segmentation or in object recognition.

C. Quantitative Evaluation

To disentangle the effect of appearance restoration from geometric supervision, we optionally apply lightweight preprocessing modules. For each RGB frame, we experiment with two classical front ends:

- **Denoising** using SDAP [17] to suppress sensor noise under extreme ISO settings;
- **Deblurring** using LaKDNet [18] to correct motion-blur artifacts.

These preprocessing modules are used to generate input variants denoted as +SDAP and +LaKDNet, respectively. Both are applied prior to feature extraction and volumetric rendering, while the remainder of the LUNA pipeline including the geometry-aware depth loss $\mathcal{L}_{\text{geo}}$ remains unchanged. This allows us to isolate how restored appearance cues influence reconstruction and segmentation quality under different degradations.

We evaluate the base LUNA model and its variants augmented with the preprocessing modules to explicitly quantify the impact of input restoration. Unless otherwise stated, “LUNA” refers to the geometry-aware model without preprocessing. Table II summarizes average performance across six Replica scenes under three levels of Gaussian noise and motion blur. Bold values indicate the best-performing method under each condition, and underlined values represent the second-best.

Under noisy conditions, our method outperforms the baseline even when the latter is combined with a denoising module, showing consistent gains in both easy and moderate settings. At strong noise levels, adding the denoiser (+SDAP) further improves stability, indicating that our geometry-aware formulation leverages structured noise patterns effectively while still benefiting from denoised inputs. In contrast, for blur degradations, the deblurring module (+LaKDNet) provides consistent improvements across all degradation levels, yielding the best overall results. This demonstrates that while LUNA alone can handle substantial appearance degradation via geometric cues, preprocessing remains complementary under extreme conditions.

D. Qualitative Results

Under strong Gaussian noise, the RGB views rendered by panoptic lifting exhibit denoising artifacts and significant loss of detail, leading to missed detections and over-smoothed semantics, as shown in Fig. 4. For instance, smaller objects near the walls and floor (e.g., the lamp) are either merged or entirely missing. In contrast, LUNA restores finer RGB details and successfully detects a larger number of instances despite severe visual corruption. This demonstrates the ability to preserve fine-grained structure and segmentation fidelity under degraded visual conditions with the integration of both input restoration and geometry-aware supervision.

Under strong motion blur, panoptic lifting struggles to maintain semantic consistency and instance boundaries, as shown in Fig. 5. The chair on the left side is only partially segmented, and several furniture edges are lost or fused into background regions. This degradation leads to fragmented instances and semantic drift, particularly around textured or occluded areas. LUNA, on the other hand, reconstructs complete object structures with clear boundary delineation, preserving object continuity and sharper depth transitions. This improvement

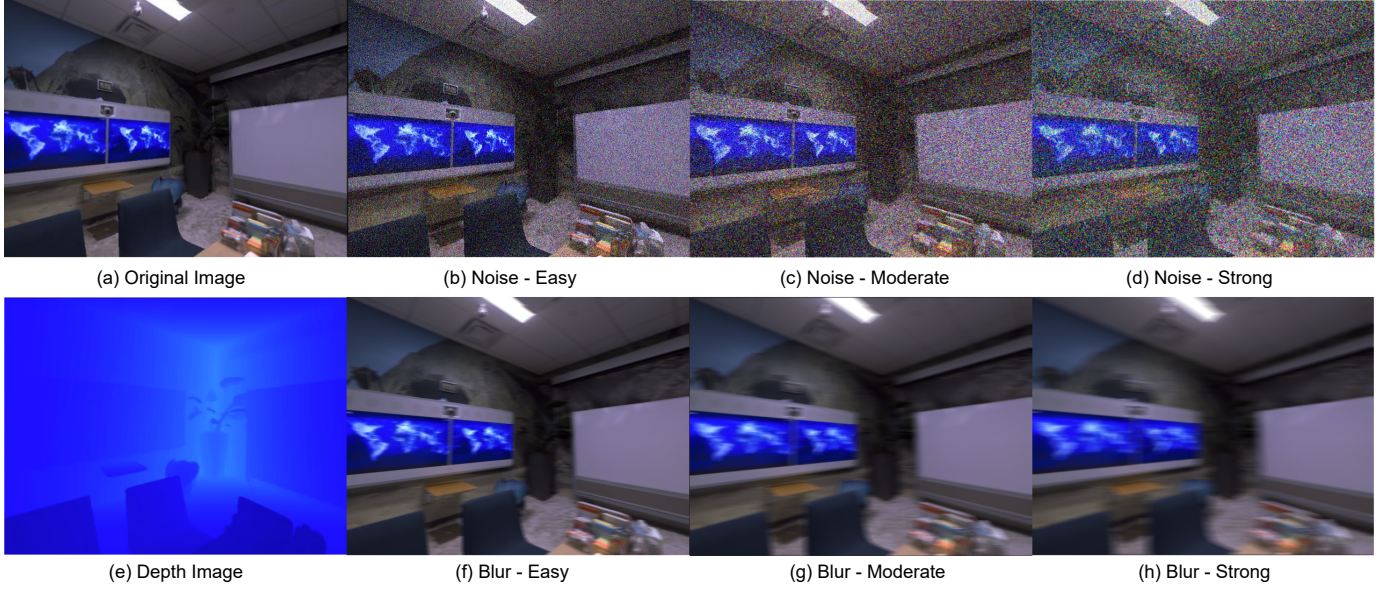


Fig. 3. Synthetic degradation examples used for evaluation. Top row: (a) clean reference RGB image, and (b–d) noisy variants generated by adding Gaussian noise with increasing standard deviations corresponding to easy, moderate, and hard levels. Bottom row: (e) ground-truth depth map (unaffected by appearance changes), and (f–h) motion-blurred RGB variants with increasing blur kernel sizes representing easy, moderate, and hard levels. These controlled degradations simulate the visual challenges commonly encountered in low-light and dynamic scenes.

TABLE II

QUANTITATIVE COMPARISON UNDER DIFFERENT DEGRADATION LEVELS. WE EVALUATE BOTH LUNA (WITHOUT PREPROCESSING) AND ITS VARIANTS WITH SDAP DENOISING AND LAKDNET DEBLURRING MODULES. RESULTS ARE AVERAGED ACROSS SIX SCENES. ARROWS ($\uparrow$) INDICATE HIGHER IS BETTER. BOLD INDICATES THE BEST RESULTS, AND UNDERLINED THE SECOND-BEST. THE TOP BLOCK SHOWS GAUSSIAN NOISE SETTINGS, AND THE BOTTOM BLOCK SHOWS MOTION BLUR SETTINGS.

Evaluation Metrics		Easy-Configuration				Moderate-Configuration				Strong-Configuration			
		PSNR $\uparrow$	SSIM $\uparrow$	mIoU $\uparrow$	PQ $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	mIoU $\uparrow$	PQ $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	mIoU $\uparrow$	PQ $\uparrow$
Noise	Panoptic Lifting	21.17	0.56	0.147	0.113	17.09	0.51	0.098	0.059	12.14	0.32	0.020	0.005
	Panoptic Lifting + SDAP	<u>24.75</u>	0.56	0.136	0.089	17.71	0.52	0.086	0.040	<u>14.72</u>	0.49	<u>0.083</u>	<u>0.038</u>
	Ours (LUNA)	23.19	0.76	0.243	0.203	<u>18.06</u>	0.65	0.174	0.184	13.21	<u>0.64</u>	0.024	0.015
	Ours (LUNA) + SDAP	25.43	<u>0.75</u>	<u>0.229</u>	<u>0.201</u>	21.49	<u>0.65</u>	<u>0.141</u>	<u>0.182</u>	16.13	0.65	0.046	0.076
Blur	Panoptic Lifting	30.17	0.92	0.215	0.195	28.19	<u>0.91</u>	0.250	0.188	27.59	0.87	0.251	0.182
	Panoptic Lifting + LaKDNet	<u>30.85</u>	0.94	<u>0.396</u>	0.267	<u>30.54</u>	0.93	0.319	0.231	28.51	0.92	<u>0.284</u>	<u>0.220</u>
	Ours (LUNA)	30.30	0.92	<u>0.392</u>	<u>0.282</u>	30.11	<u>0.91</u>	<u>0.340</u>	<u>0.279</u>	<u>29.19</u>	0.88	0.255	0.195
	Ours (LUNA) + LaKDNet	31.43	0.95	0.465	0.343	30.75	0.93	0.378	0.329	30.13	<u>0.90</u>	0.289	0.316

arises from gradient-based geometric supervision and adaptive loss balancing via FAMO, which guide the model to rely on consistent geometric cues.

E. Ablation on Depth Supervision Losses.

For completeness, we also evaluate two additional depth losses commonly used in geometry supervision. The *scale-invariant loss* [25] measures relative depth error independent of global scale:

$$\mathcal{L}_{\text{SI}} = \frac{1}{n} \sum_i d_i^2 - \frac{\lambda}{n^2} \left(\sum_i d_i \right)^2, \quad (11)$$

where $d_i = \log \hat{z}_i - \log z_i$ denotes the per-pixel log depth difference, and $\lambda = 0.5$ balances absolute and relative terms.

The *normal-based loss* [26] enforces surface-level consistency by aligning predicted and ground-truth normals:

$$\mathcal{L}_{\text{norm}} = \frac{1}{n} \sum_i \left(1 - \mathbf{n}_i^\top \hat{\mathbf{n}}_i \right), \quad (12)$$

where $\mathbf{n}_i$ and $\hat{\mathbf{n}}_i$ denote the ground-truth and predicted surface normals at pixel i , respectively.

Table III presents an ablation study comparing the four loss components used within our geometry-aware depth supervision module: the $\mathcal{L}_1$ loss, depth gradient loss, normal-based loss, and scale-invariant (SI) loss. Each component is applied individually while keeping the rest of the training pipeline fixed. Results are reported as performance differences relative to our full LUNA model, which combines the $\mathcal{L}_1$ and gradient losses. Compared to LUNA, most single-loss variants show reduced performance across all metrics (negative gains), indicating that the combined formulation

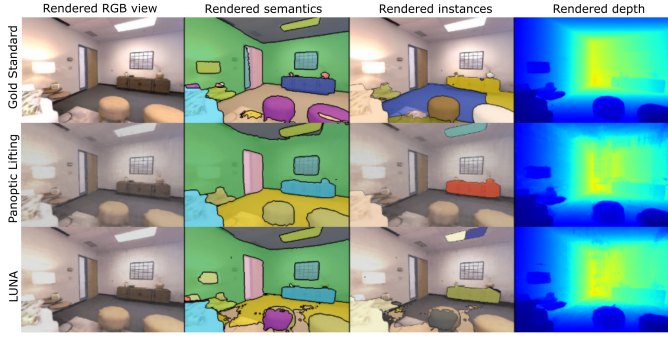


Fig. 4. Qualitative comparison of panoptic segmentation under strong Gaussian noise. Top to bottom: ground-truth, panoptic lifting, and our method LUNA. Panoptic lifting fails to recover fine semantic details, missing smaller objects such as lamps and side-table accessories, and produces over-smoothed depth maps with weakened geometric boundaries. In contrast, LUNA restores clearer RGB details, preserves spatial layout, and segments more instances even under severe noise, demonstrating robustness from geometry-aware depth supervision.



Fig. 5. Qualitative comparison of panoptic segmentation under strong motion blur. Top to bottom: ground truth, panoptic lifting, and our method LUNA. Panoptic lifting exhibits incomplete segmentation and semantic drift, with only partial detection of key objects (e.g., the chair). LUNA reconstructs complete object boundaries and sharper depth edges, demonstrating robustness from geometry-aware depth supervision and adaptive multitask optimization.

provides more balanced supervision. Among the individual variants, the SI loss yields the highest SSIM, highlighting its strength in capturing structural similarity, while the normal-based loss achieves the best SQ, reflecting improved surface-level segmentation quality. We also observe that the $\mathcal{L}_1$ loss excels in appearance reconstruction (PSNR), and the gradient loss better supports downstream segmentation metrics (mIoU, PQ, RQ). Overall, combining $\mathcal{L}_1$ and gradient losses in LUNA consistently achieves the most stable and superior performance across both reconstruction and segmentation tasks.

V. DISCUSSION

A. Recommendations

Based on the empirical trends observed in our quantitative and qualitative analyses (Sec. IV-C, Sec. IV-D, Sec. IV-E), we outline the following recommendations for deploying 3D scene understanding systems in visually degraded robotic environments:

- **Static or slow-moving platforms:** In settings with minimal camera motion, longer exposure times improve the

TABLE III
ABLATION OF DEPTH SUPERVISION LOSS VARIANTS ON *Noise-Moderate* IMAGES, AVERAGED OVER *room_0* AND *office_0*. WE REPORT PERFORMANCE GAINS RELATIVE TO OUR PROPOSED METHOD. **BOLD** INDICATES THE BEST PERFORMANCE AMONG SINGLE-LOSS VARIANTS.

Method	PSNR $\uparrow$	SSIM $\uparrow$	mIoU $\uparrow$	PQ $\uparrow$	RQ $\uparrow$	SQ $\uparrow$
$\mathcal{L}_1$	-0.11	-0.14	-0.002	-0.031	-0.009	-0.031
Gradient	-0.51	-0.19	-0.008	-0.027	+0.001	-0.013
Normal	-0.12	-0.38	+0.012	-0.037	-0.043	-0.011
SI	-1.63	-0.13	-0.064	-0.153	-0.196	-0.130

SNR by capturing more photons, reducing noise without introducing significant blur. Consistent with our results in Table II, deblurring (+LaKDNet) yielded larger gains than denoising (+SDAP) under mild motion, suggesting that blur correction is more beneficial than noise suppression for stationary or low-speed inspection tasks.

- **Dynamic environments:** For mobile robots, where motion blur dominates, higher ISO and shorter exposure are often unavoidable. In these cases, maintaining a higher depth frame rate than RGB ensures geometric consistency for volumetric optimization, aligning with our observation that LUNA’s geometry-aware depth loss preserved scene structure despite appearance degradation. Thus, depth synchronization and high-frequency geometric updates are recommended for drones or fast-moving platforms.
- **Highly degraded conditions:** Under severe noise or blur, geometry-aware supervision consistently delivered the largest performance gains, as shown in Table II. Denoising sometimes introduced artifacts in textureless regions, while depth supervision remained stable. Empirically, MLP-based depth concatenation performed better with denoised inputs, whereas direct geometry supervision proved more effective under blur, supporting the use of structured depth cues when appearance information is unreliable.
- **Computational considerations:** As observed in our ablations (Table III), depth-augmented semantic prediction provides a favorable trade-off between accuracy and efficiency, achieving notable robustness improvements under moderate degradations without requiring the full geometry-aware loss pipeline. This lightweight configuration is therefore recommended for resource-constrained robotic platforms.

B. Limitations

While LUNA improves robustness under low-light degradations, its performance remains tied to the underlying panoptic segmentation backbone. In particular, Mask2Former can become a bottleneck under severe noise or blur due to its reliance on appearance cues, as shown in Fig. 6. Under heavy noise, objects with similar spatial extents or textures, such as a whiteboard being misclassified as a television illustrates how appearance corruption leads to semantic confusion and label bias. The resulting segmentation shows fragmented and

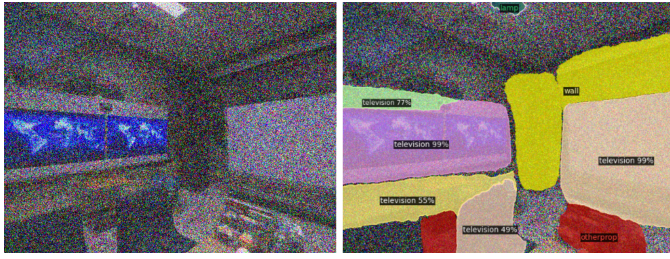


Fig. 6. Failure case under severe low-light degradation. Heavy noise severely corrupts appearance cues, resulting in misclassified instances and highlighting the limitations of RGB-based segmentation under extreme conditions.

mislabelled regions, highlighting the limits of RGB-based perception when structure cues are overwhelmed by noise.

Our framework also assumes a fixed semantic vocabulary, which may restrict open-world generalization. It further depends on high-quality ground-truth depth, whereas real-world LiDAR or stereo data often contain noise and sparsity. Finally, LUNA has so far been validated only on synthetic datasets, and real-world deployment may introduce additional domain shifts and calibration challenges.

VI. CONCLUSION

We introduced LUNA, a geometry-aware extension to panoptic lifting that integrates depth supervision and panoptic segmentation under challenging image degradations. By unifying depth accuracy and gradient smoothness into a geometric supervision signal, and jointly optimizing photometric, semantic, instance, and geometric objectives via FAMO, LUNA outperforms both the panoptic lifting baseline and restoration-augmented pipelines across reconstruction and segmentation metrics. These results highlight the value of geometry-driven supervision and adaptive loss weighting for achieving robust 3D perception. Future work will investigate richer adaptive scheduling, realistic sensor noise, and on-platform deployment in unstructured real-world environments.

REFERENCES

- [1] X. Liu, Z. Wu, A. Li, F.-A. Vasluianu, Y. Zhang, S. Gu, L. Zhang, C. Zhu, R. Timofte, Z. Jin, *et al.*, “NTIRE 2024 Challenge on Low Light Image Enhancement: Methods and Results,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6571–6594.
- [2] G. Wang, L. Pan, S. Peng, S. Liu, C. Xu, Y. Miao, W. Zhan, M. Tomizuka, M. Pollefeys, and H. Wang, “NeRF in Robotics: A Survey,” *arXiv preprint arXiv:2405.01333*, 2024.
- [3] B. Fei, J. Xu, R. Zhang, Q. Zhou, W. Yang, and Y. He, “3D Gaussian Splatting as New Era: A Survey,” *IEEE Transactions on Visualization and Computer Graphics*, 2024.
- [4] Y. Siddiqui, L. Porzi, S. R. Buló, N. Müller, M. Nießner, A. Dai, and P. Kotschieder, “Panoptic Lifting for 3D Scene Understanding with Neural Fields,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9043–9052.
- [5] H. Chen, K. Blomqvist, F. Milano, and R. Siegwart, “Panoptic Vision-Language Feature Fields,” *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2144–2151, 2024.
- [6] H. Zhai, H. Li, Z. Li, X. Pan, Y. He, and G. Zhang, “PanoGS: Gaussian-based Panoptic Segmentation for 3D Open Vocabulary Scene Understanding,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 114–14 124.
- [7] A. Ravendran, M. Bryson, and D. G. Dansereau, “Burst Imaging for Light-Constrained Structure-from-Motion,” *Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 1040–1047, 2021.
- [8] H. Wang, X. Xu, K. Xu, and R. W. Lau, “Lighting up NeRF via Unsupervised Decomposition and Enhancement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 632–12 641.
- [9] B. Mildenhall, P. Hedman, R. Martin-Brualla, P. P. Srinivasan, and J. T. Barron, “NeRF in the dark: High dynamic range view synthesis from noisy raw images,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 190–16 199.
- [10] Z. Liu, Q. Meng, D. Lu, G. Li, D. Guo, D. Liu, C. Liu, and Z. Yang, “GloNeRF: Boosting NeRF Capabilities and Multi-View Consistency in Low-Light Environments,” *Computers & Graphics*, vol. 128, p. 104209, 2025.
- [11] B. Liu, Y. Feng, P. Stone, and Q. Liu, “FAMO: Fast Adaptive Multitask Optimization,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 57 226–57 243, 2023.
- [12] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijnmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, *et al.*, “The Replica Dataset: A Digital Replica of Indoor Spaces,” *arXiv preprint arXiv:1906.05797*, 2019.
- [13] Y. Cai, Z. Xiao, Y. Liang, M. Qin, Y. Zhang, X. Yang, Y. Liu, and A. L. Yuille, “HDR-GS: Efficient High Dynamic Range Novel View Synthesis at 1000x Speed via Gaussian Splatting,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 68 453–68 471, 2024.
- [14] S. Ye, Z.-H. Dong, Y. Hu, Y.-H. Wen, and Y.-J. Liu, “Gaussian in the Dark: Real-Time View Synthesis From Inconsistent Dark Images Using Gaussian Splatting,” in *Computer Graphics Forum*, vol. 43, no. 7. Wiley Online Library, 2024, p. e15213.
- [15] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, “Burst Photography for High Dynamic Range and Low-Light Imaging on Mobile Cameras,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.
- [16] O. Liba, K. Murthy, Y.-T. Tsai, T. Brooks, T. Xue, N. Karnad, Q. He, J. T. Barron, D. Sharlet, R. Geiss, *et al.*, “Handheld Mobile Photography in Very Low Light,” *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–16, 2019.
- [17] Y. Pan, X. Liu, X. Liao, Y. Cao, and C. Ren, “Random Sub-Samples Generation for Self-Supervised Real Image Denoising,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 150–12 159.
- [18] L. Ruan, M. Berman, H.-P. Seidel, K. Myszkowski, and B. Chen, “Revisiting Image Deblurring with an Efficient ConvNet,” *arXiv preprint arXiv:2302.02234*, 2023.
- [19] A. Ravendran, M. Bryson, and D. G. Dansereau, “BuFF: Burst Feature Finder for Light-Constrained 3D Reconstruction,” *Robotics and Automation Letters (RA-L)*, vol. 8, no. 12, pp. 8438–8445, 2023.
- [20] G. Narita, T. Seno, T. Ishikawa, and Y. Kaji, “PanopticFusion: Online Volumetric Semantic Mapping at the Level of Stuff and Things,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2019, pp. 4205–4212.
- [21] L. Schmid, J. Delmerico, J. L. Schönberger, J. Nieto, M. Pollefeys, R. Siegwart, and C. Cadena, “Panoptic Multi-TSDFs: A Flexible Representation for Online Multi-Resolution Volumetric Mapping and Long-term Dynamic Scene Consistency,” in *International Conference on Robotics and Automation*. IEEE, 2022, pp. 8018–8024.
- [22] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing, “Mask2Former for Video Instance Segmentation,” *arXiv preprint arXiv:2112.10764*, 2021.
- [23] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, “TensorRF: Tensorial Radiance Fields,” in *European Conference on Computer Vision*. Springer, 2022, pp. 333–350.
- [24] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, “Panoptic Segmentation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9404–9413.
- [25] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *Advances in neural information processing systems*, vol. 27, 2014.
- [26] W. Yin, Y. Liu, C. Shen, and Y. Yan, “Enforcing geometric constraints of virtual normal for depth prediction,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5684–5693.