

SAFARI: Semantic–Depth Adaptive SLAM for Visually Challenging Robotic Environments

Lyes Saad Saoud

`lyes.saoud@mbzuai.ac.ae`

Khalifa University

Reza Ghorbani

University of Hawai'i at Mānoa

Irfan Hussain

Khalifa University

Research Article

Keywords: semantic SLAM, adaptive SLAM, underwater robotics, vision–language models, monocular depth estimation, Gaussian mapping

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11198179/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SAFARI: Semantic–Depth Adaptive SLAM for Visually Challenging Robotic Environments

Lyes Saad Saoud^{1,*} Reza Ghorbani² Irfan Hussain¹

¹Khalifa University, Abu Dhabi, United Arab Emirates

²University of Hawai‘i at Mānoa, Honolulu, HI, USA

*Corresponding author: Lyes Saad Saoud

Email: lyes.saoud@mbzuai.ac.ae

Abstract

Robust simultaneous localization and mapping (SLAM) remains difficult when the visual appearance of a scene changes faster than a fixed SLAM configuration can accommodate. Low texture, reflections, turbidity, partial occlusion, and repetitive structures can simultaneously degrade feature extraction, monocular depth prediction, and map optimization. This paper presents SAFARI, a semantic–depth adaptive SLAM framework that couples dense monocular depth, vision–language scene interpretation, and constrained online parameter adaptation. Depth Anything V2 supplies dense geometric cues, CLIP produces scene-level semantic evidence, and a lightweight language-model controller converts compact perception and optimization statistics into bounded updates of mapping parameters such as keyframe density, optimization iterations, and pruning behavior. A semantic–depth confidence module attenuates unreliable depth in semantically identified difficult regions before the observations are incorporated into the map. The framework is evaluated on the Replica indoor benchmark and on a marine-pool sequence acquired with an underwater remotely operated vehicle (KUMP). On the reported Replica ablation, semantic masking increases reconstruction PSNR from 20.18 to 34.25 dB and reduces depth L1 error from 0.23 to approximately zero at the evaluated frame, while yielding an IoU of 0.8916. On the underwater sequence, the complete pipeline achieves an ATE RMSE of 41.79 cm, RGB PSNR of 19.12 dB, and depth L1 error of 7.79 cm over the evaluated 200-frame sequence. These experiments support the feasibility of context-conditioned SLAM adaptation, while also exposing current limitations in long-horizon validation, underwater semantic segmentation, and computational efficiency.

Keywords: semantic SLAM; adaptive SLAM; underwater robotics; vision–language models; monocular depth estimation; Gaussian mapping.

1 Introduction

Simultaneous localization and mapping (SLAM) is a core capability for mobile robots operating without a complete prior map. Mature geometric systems can achieve high accuracy when image formation, texture, motion, and illumination remain close to their design assumptions. In practical robotic deployment, however, the same platform may move between textured and texture-poor regions, encounter reflections or repetitive structures, or experience severe appearance changes. Underwater operation intensifies these difficulties through turbidity, wavelength-dependent attenuation, artificial-light artifacts, suspended particles, and non-Lambertian reflections. Consequently, a configuration that is effective in one portion of a trajectory may become brittle in another.

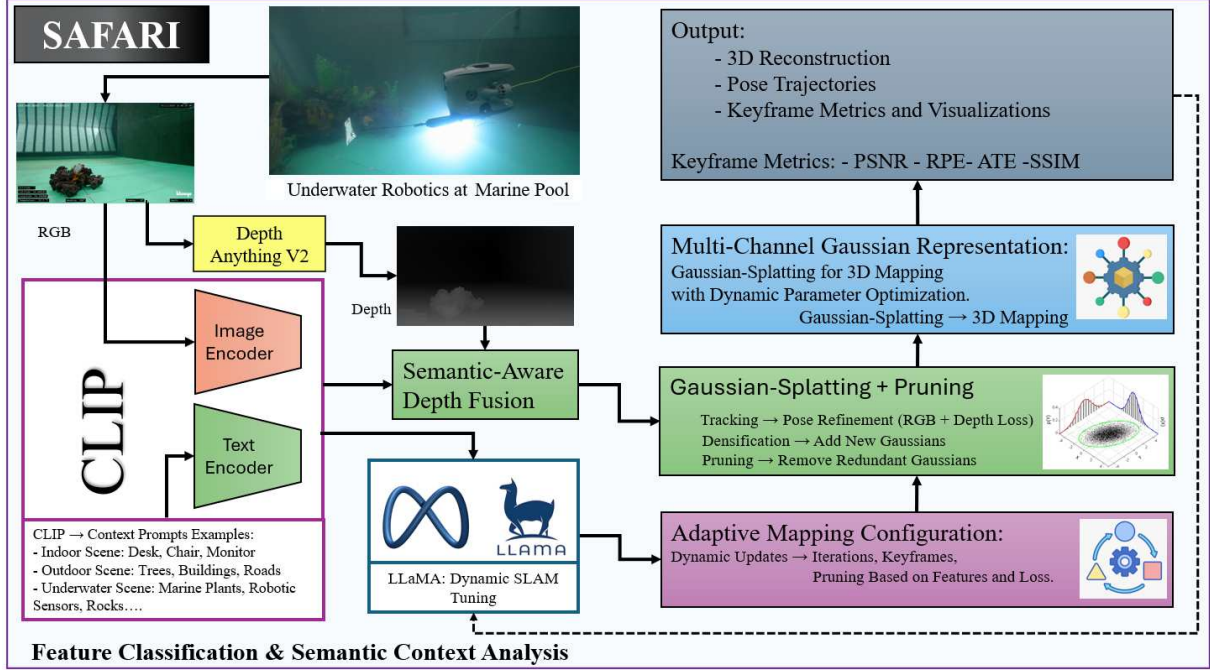


Figure 1: Overview of SAFARI. Dense depth and vision–language semantic evidence are converted into confidence-weighted geometric observations and compact scene descriptors. A bounded adaptive controller then modifies selected mapping parameters rather than replacing the underlying geometric estimator. This separation keeps the learned reasoning modules in a supervisory role while pose and map estimation remain governed by geometric optimization.

A large body of work has strengthened SLAM using learned perception. Semantic SLAM systems use object or region labels to suppress dynamic features, improve data association, or enrich the map [7–9]. Learned depth models provide dense geometry when direct depth sensing is unavailable, and recent neural or Gaussian representations support dense reconstruction and differentiable map optimization [10, 11, 23]. These developments improve individual perception or representation components, but most deployed systems still rely on manually selected thresholds, optimization budgets, keyframe policies, and pruning rules. This creates a second failure mode: even if learned perception correctly recognizes that the scene has changed, the SLAM back end may continue operating with parameters selected for a different regime.

SAFARI addresses this gap by treating semantic reasoning as a supervisory signal for *bounded adaptation*. The framework does not ask a language model to estimate pose or generate a map. Instead, geometry remains responsible for tracking and reconstruction, while learned models describe scene context and the controller selects among safe, predefined parameter changes. This division is important for robotics: a high-level model can help identify that a scene is reflective, cluttered, or low-texture, but its output should not bypass the metric constraints of the estimator.

The framework combines three information channels. First, Depth Anything V2 produces dense monocular depth for regions where sparse keypoints are insufficient [23]. Second, CLIP maps the image into a compact semantic description using zero-shot visual–text similarity [14]. Third, a lightweight LLaMA-family controller receives scene descriptors and mapping statistics and proposes parameter updates from a constrained set [18]. A semantic–depth confidence stage modifies the influence of depth measurements according to scene evidence before map optimization.

The work is evaluated in two complementary settings. Replica provides controlled indoor geometry and enables an ablation of semantic masking, while the KUMP marine-pool sequence tests the pipeline under reflections, reduced contrast, and underwater appearance changes. The goal of the present evaluation is not to claim universal superiority to established SLAM systems; rather, it is to determine whether semantic context can be used to modify the behavior of a mapping system in a measurable and interpretable way.

The main contributions are:

- **A context-conditioned SLAM architecture** in which dense depth, zero-shot visual semantics, and map statistics are integrated without replacing geometric pose estimation.
- **A bounded adaptive-control mechanism** that translates semantic and optimization evidence into a limited set of parameter updates, including keyframe density, optimization effort, and pruning behavior.
- **A semantic–depth confidence formulation** that reduces the contribution of depth estimates in scene regions identified as unreliable or reflection-prone.
- **Cross-domain evaluation** on the Replica benchmark and an underwater ROV sequence, including a semantic-mask ablation and quantitative trajectory, reconstruction, depth, and semantic measurements.

2 Related Work

2.1 Learning-Enhanced and Semantic SLAM

Deep models have been integrated into SLAM primarily for feature extraction, dynamic-object rejection, semantic mapping, and learned correspondence. YOLO-based pipelines, for example, use object detections to exclude unstable regions or to improve scene interpretation [19, 21]. Dynamic-aware methods combine segmentation or tracking with geometric estimation so that independently moving objects contribute less to pose estimation [8, 24]. Vision transformers and large-scale visual encoders further improve global context representation in scenes where local keypoints are ambiguous [2, 6].

These methods demonstrate that semantics can improve geometric estimation, but semantic evidence is commonly consumed as a mask or map label. SAFARI differs by also using scene semantics to influence selected optimization hyperparameters. The distinction is between *semantic perception* and *semantic supervision of estimator behavior*.

2.2 Dense Depth and Neural Scene Representations

RGB-D and stereo sensors provide direct or geometric depth, whereas monocular systems infer depth from appearance and learned priors [4, 17]. Foundation-scale monocular models such as Depth Anything improve dense prediction robustness across diverse visual domains [23]. In parallel, neural implicit and Gaussian scene representations have enabled dense online mapping, differentiable rendering, and joint camera–scene optimization [10, 11].

Dense learned depth is useful in weakly textured regions, but monocular estimates remain uncertain, especially under domain shift. Underwater images are a particularly severe example because color attenuation, scattering, and specular reflections can violate assumptions learned from terrestrial training data. SAFARI therefore treats dense depth as a confidence-weighted cue rather than as ground truth.

Table 1: Positioning of SAFARI relative to representative learning-enhanced SLAM directions. Reported benefits in prior work are summarized qualitatively because experimental protocols differ across studies.

Direction	Learned component	Primary role	Relation to SAFARI
Dynamic/semantic SLAM [8, 21, 24]	detector / segmenter	reject dynamic regions; enrich maps	SAFARI also uses semantics to supervise parameter adaptation
Learned depth SLAM [17, 23]	monocular depth network	dense geometric cue	SAFARI confidence-weights depth using semantic evidence
Gaussian/neural mapping [10, 11]	differentiable scene representation	dense reconstruction and optimization	SAFARI uses a Gaussian-style map representation with adaptive optimization
Underwater enhancement [3, 12, 20]	restoration/enhancement network	improve visual observability	SAFARI combines preprocessing with scene-conditioned mapping decisions
SAFARI	depth foundation model + VLM + constrained LLM controller	geometry, context, and bounded adaptation	supervisory reasoning modifies selected SLAM parameters while preserving geometric estimation

2.3 Underwater SLAM and Visually Degraded Environments

Underwater localization has long relied on combinations of visual, inertial, acoustic, and filtering methods [5, 13]. Image enhancement and learned restoration have been explored to improve visibility before feature extraction [3, 12, 20], while self-supervised approaches seek features that remain stable under underwater appearance variation [22]. Datasets such as SVIn2 and MIMIR-UW expose algorithms to low illumination, haze, reflections, or backscatter [1, 15].

The remaining challenge is not only perception quality but configuration robustness. An underwater robot may traverse clear and turbid areas within a single mission. Fixed thresholds and fixed optimization budgets can therefore be suboptimal even when the underlying SLAM method is unchanged. SAFARI focuses specifically on this regime-dependent adaptation problem.

3 SAFARI Framework

Let the input sequence be $\mathcal{I} = \{I_t\}_{t=1}^N$ with camera intrinsics K . At time t , SAFARI maintains a camera pose T_t , a map $\mathcal{M}_t$, and a parameter vector θ_t containing only parameters that are allowed to change online. The processing loop is

$$(D_t, z_t, q_t) = \Phi(I_t, \mathcal{M}_{t-1}), \quad (1)$$

where D_t is predicted depth, z_t is semantic context, and q_t denotes compact estimator statistics. A bounded policy π proposes

$$\Delta\theta_t = \Pi_{\mathcal{A}}(\pi(z_t, q_t)), \quad \theta_t = \text{clip}(\theta_{t-1} + \Delta\theta_t, \theta_{\min}, \theta_{\max}), \quad (2)$$

where $\Pi_{\mathcal{A}}$ projects the controller output onto a predefined action set and $\text{clip}(\cdot)$ enforces parameter-specific limits. This formulation explicitly prevents unrestricted language-model output from directly controlling the estimator.

3.1 Dense Depth Estimation

For each frame, Depth Anything V2 predicts a dense relative-depth field

$$\hat{D}_t = f_{\text{depth}}(I_t). \quad (3)$$

The prediction is used as a geometric prior during reconstruction and tracking. Because monocular depth can be unreliable under domain shift, particularly underwater, the raw estimate is not treated as a metric sensor measurement. Instead, it is modulated by a confidence map derived from semantic context and visual reliability indicators.

3.2 Semantic Scene Interpretation

CLIP embeds the image and a set of textual scene prompts into a shared representation. For a prompt set $\mathcal{P} = \{p_j\}$, the normalized confidence is

$$c_{t,j} = \frac{\exp(\tau^{-1} \text{sim}(e_I(I_t), e_T(p_j)))}{\sum_k \exp(\tau^{-1} \text{sim}(e_I(I_t), e_T(p_k)))}, \quad (4)$$

where e_I and e_T are the image and text encoders and τ is the temperature. The top semantic labels form z_t . Prompts describe scene properties and structures relevant to mapping, for example reflective surfaces, repetitive textures, vegetation-like regions, walls, desks, or underwater objects.

3.3 Semantic–Depth Confidence Fusion

The original draft multiplied depth directly by semantic confidence. For a SLAM system, this can incorrectly alter scale. We instead formulate semantic information as a *measurement weight*. Let $w_t(x) \in [w_{\min}, 1]$ represent confidence at pixel x . The depth residual is

$$r_D(x) = w_t(x)(D_t^{\text{ref}}(x) - \hat{D}_t(x)), \quad (5)$$

and the weighted depth term is

$$L_D = \sum_{x \in \Omega_t} \rho(r_D(x)), \quad (6)$$

where ρ is a robust penalty. Semantic classes or scene cues associated with unreliable geometry, such as strongly reflective regions, receive lower w_t , whereas stable planar structures can retain higher confidence. This preserves the numerical depth prediction while changing only its influence on optimization.

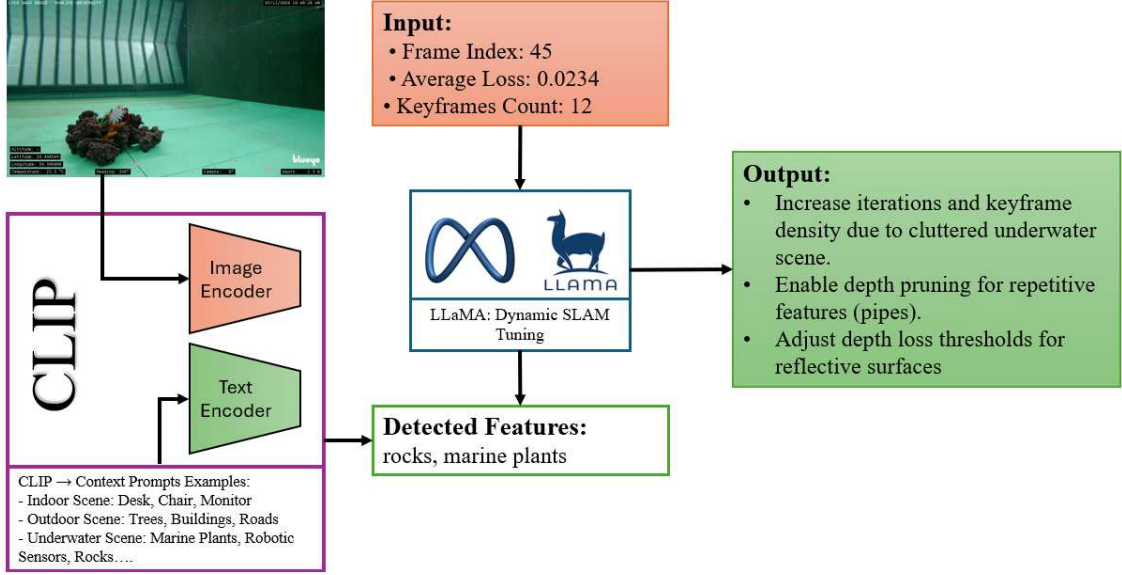


Figure 2: Semantic supervision pathway. CLIP produces compact scene evidence; the controller receives this evidence together with loss and keyframe statistics and selects from bounded mapping actions.

3.4 Mapping and Tracking Objective

The map is represented by Gaussian primitives with center μ_i , scale/covariance parameters, opacity α_i , appearance, and optional semantic attributes. A simplified isotropic form is

$$G_i(x) = \alpha_i \exp \left[-\frac{\|x - \mu_i\|^2}{2r_i^2} \right]. \quad (7)$$

Differentiable rendering provides predicted color $\hat{C}_t$, depth $\hat{D}_t$, and silhouette $\hat{S}_t$. The optimization objective is

$$L = \lambda_C L_C + \lambda_D L_D + \lambda_S L_S, \quad (8)$$

with photometric, confidence-weighted depth, and semantic/silhouette terms. In the reported configuration, $\lambda_D = 1.0$, $\lambda_C = 0.5$, and $\lambda_S = 0.05$, with a silhouette visibility threshold of $T_{sil} = 0.99$ and camera-pose learning rate 2×10^{-3} .

3.5 Bounded Adaptive Controller

The controller receives a compact text representation of (z_t, q_t) , where q_t includes current loss, recent loss trend, keyframe count, and selected map statistics. Rather than accepting arbitrary generated instructions, SAFARI parses the model response into a small action vocabulary such as

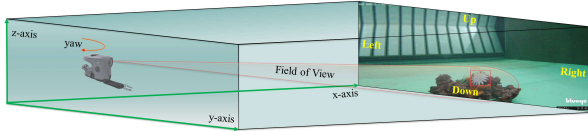
$$\mathcal{A} = \{\text{increase_iterations, decrease_iterations, add_keyframe, prune, hold}\}. \quad (9)$$

Each action maps to a deterministic parameter increment and is clipped to an allowed range as in Eq. (2). This design is more reproducible and safer than directly executing free-form recommendations.

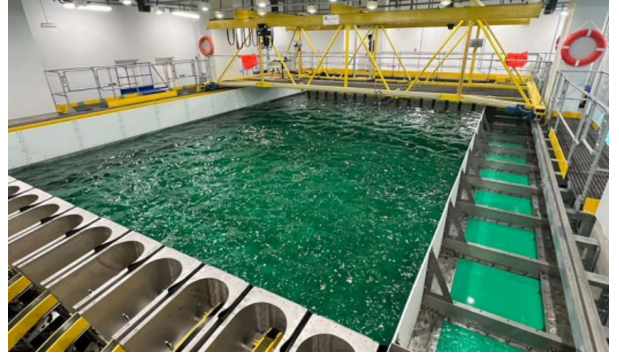
Algorithm 1 SAFARI online processing loop

Require: frames I_t , intrinsics K , initial map $\mathcal{M}_0$, bounds $(\theta_{\min}, \theta_{\max})$

- 1: **for** $t = 1, \dots, N$ **do**
 - 2: predict dense depth $\hat{D}_t$
 - 3: infer semantic context z_t with CLIP
 - 4: compute semantic confidence weights w_t
 - 5: track camera and update the map using Eq. (8)
 - 6: summarize optimization statistics q_t
 - 7: query controller and project output to $\mathcal{A}$
 - 8: update bounded parameters using Eq. (2)
 - 9: **end for**
 - 10: **return** trajectory $\{T_t\}$ and map $\mathcal{M}_N$
-



(a) ROV field of view.



(b) Marine-pool environment.

Figure 3: KUMP underwater acquisition setting.

4 Experimental Methodology

4.1 Hardware and Software

Experiments were run on a laptop-class platform with an NVIDIA GeForce RTX 3050 Ti GPU and CUDA 11.8. The software stack used Python 3.11.5 and PyTorch 2.2.1. The current implementation prioritizes research transparency over embedded optimization; consequently, runtime results should be interpreted as prototype performance rather than as proof of high-rate onboard deployment.

4.2 Datasets

Replica. Replica provides photorealistic indoor scenes with ground-truth geometry and camera information [16]. Room0 is used for the reported semantic-mask ablation because it contains structured surfaces and objects that permit direct comparison of reconstruction with and without semantic guidance.

KUMP underwater sequence. The second evaluation uses a marine-pool sequence collected with a FIFISH ROV. The setting contains reflective surfaces, artificial lighting, low-contrast regions, and underwater appearance variations. The evaluated sequence contains 200 frames in the current experiments. The dataset is used as a difficult deployment case rather than as a replacement for established large-scale underwater benchmarks.

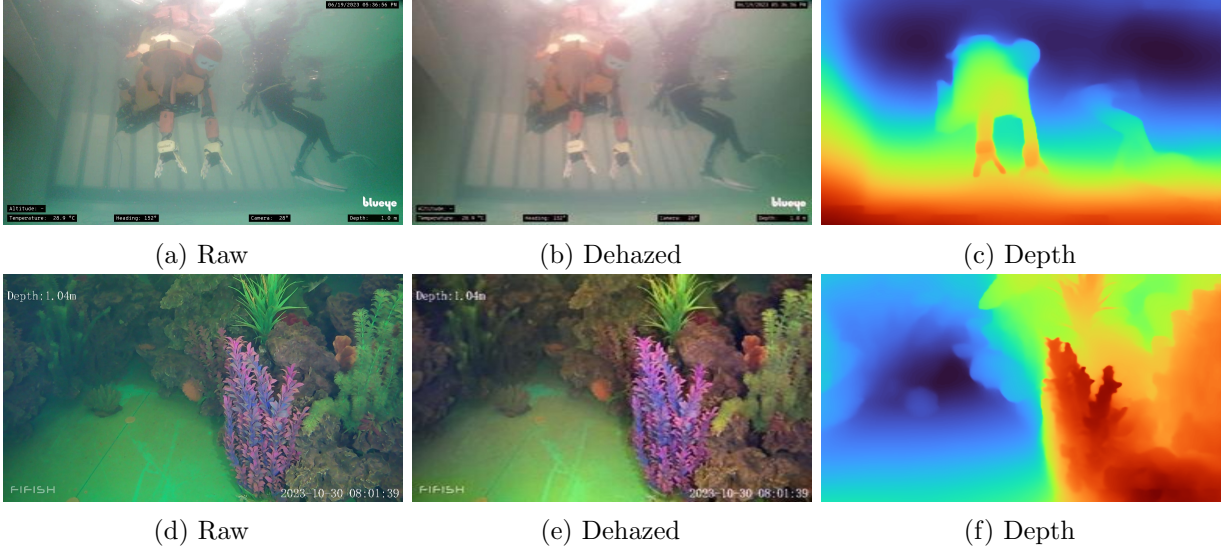


Figure 4: Examples of underwater preprocessing and monocular depth prediction.

4.3 Preprocessing

Underwater images are optionally dehazed before depth prediction. Figure 4 shows representative raw, enhanced, and depth-prediction triplets. This preprocessing is held fixed during the reported sequence so that changes in mapping behavior are attributable to the adaptive module rather than to manual retuning between frames.

4.4 Metrics and Evaluation Protocol

Trajectory accuracy is measured with Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) when corresponding reference poses are available. Reconstruction is assessed with PSNR, depth L1 error, and semantic overlap (IoU/mIoU where applicable). Runtime is measured separately because adaptive reasoning can introduce overhead that is not reflected in reconstruction metrics.

The current study contains two distinct evaluations: (i) a controlled Replica ablation of semantic masking, and (ii) an end-to-end underwater sequence evaluation. Results from these two protocols are reported separately to avoid implying that a metric from one dataset is directly comparable with a metric from the other.

5 Results

5.1 Replica Semantic-Mask Ablation

Figures 5 and 6 compare the same Replica example with and without semantic masking. The semantic configuration achieves PSNR 34.25 dB, depth L1 approximately 0.00 in the reported rendering, and IoU 0.8916. Without semantic masks, PSNR is 20.18 dB and depth L1 is 0.23. The result demonstrates a clear effect in this controlled example, but it should not be interpreted as a dataset-wide percentage improvement because the present manuscript does not report a full multi-sequence statistical ablation.

CLIP also provides interpretable scene descriptors. Table 2 lists the highest-confidence labels observed in the reported Room0 analysis. The values are used as controller context rather than as

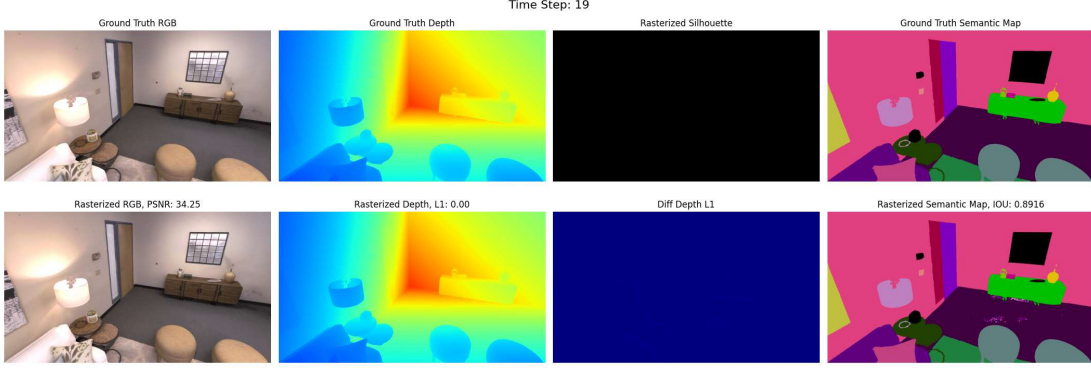


Figure 5: Replica Room0 example with semantic masks: reference RGB/depth, rendered RGB/depth, silhouette, and semantic output.

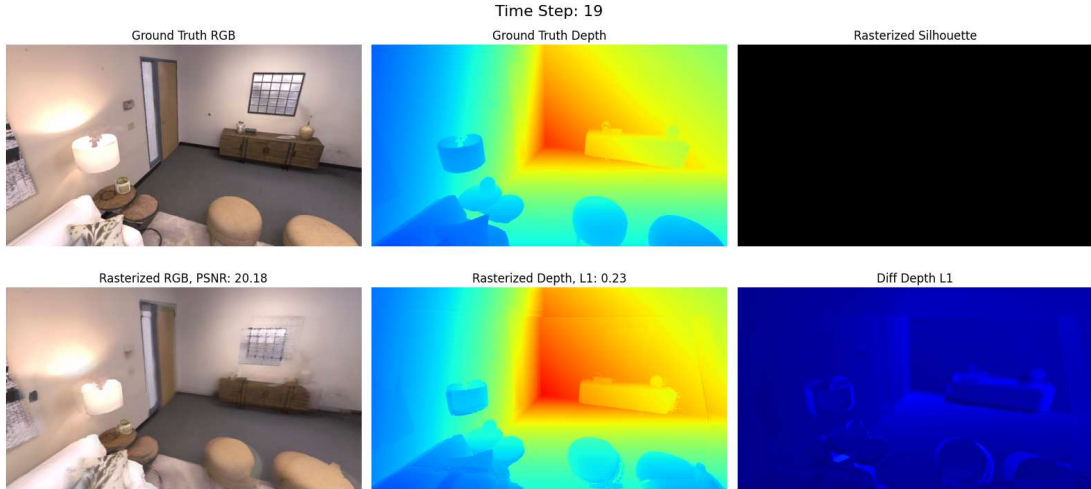


Figure 6: Replica Room0 example without semantic masks.

semantic ground truth.

5.2 Underwater Sequence

Figure 7 shows the estimated trajectory in multiple projections. Across the evaluated 200-frame sequence, the complete pipeline reports an ATE RMSE of 41.79 cm, RGB PSNR of 19.12 dB, and depth L1 error of 7.79 cm. These values quantify the current prototype under a visually difficult underwater sequence; they are not claimed to establish superiority over ORB-SLAM3, DROID-SLAM, or other baselines because matched baseline runs are not included in the present evidence package.

5.3 Semantic Evidence and Adaptive Behavior

Figure 8 shows temporal confidence for underwater semantic prompts, while Fig. 9 summarizes average confidence. The controller uses these values jointly with map statistics; individual semantic labels are not sufficient to trigger an action in isolation. This reduces sensitivity to one spurious label and makes the adaptation logic easier to audit.

Table 2: Representative CLIP detections in Replica Room0.

Feature	Confidence	Time index
Carpets	0.470	18
Desks	0.472	9
Bookshelves	0.091	19
Walls	0.084	8
Chairs	0.048	7

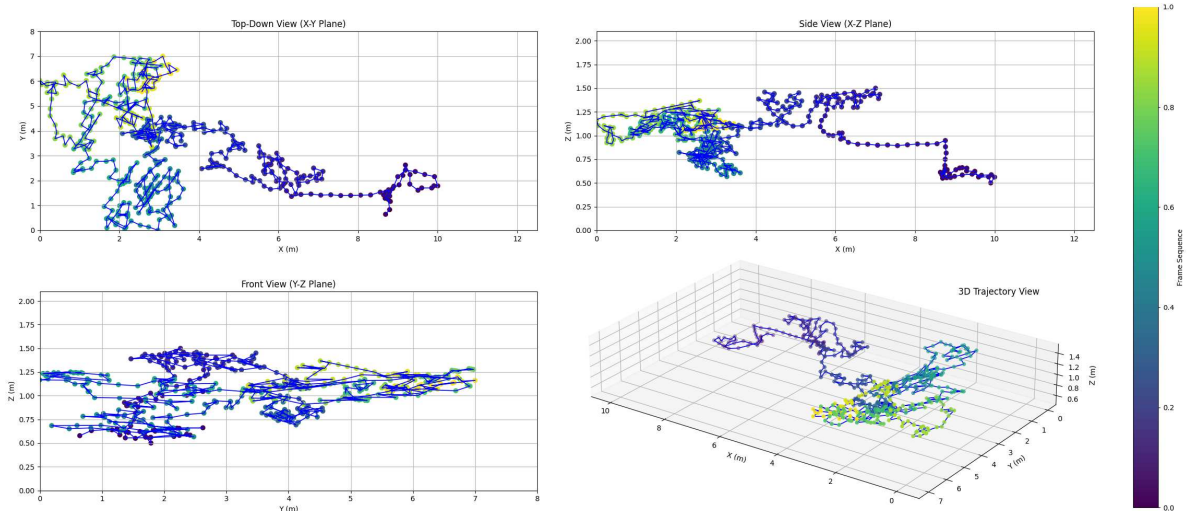


Figure 7: Estimated trajectory for the underwater experiment shown in multiple projections.

The heatmap in Fig. 10 visualizes the relationship between detected scene evidence and adaptive recommendations. The key point is not that the language model “solves” SLAM; rather, it provides a compact mechanism for mapping heterogeneous cues into a constrained vocabulary of decisions.

Figure 11 compares depth statistics before and after confidence-guided fusion. In the available experiment, the mean reported depth changes from 1.09 m to 1.20 m and frame-to-frame variation is reduced. Because the sequence does not provide dense independent metric ground truth at every pixel, the reduced variance should be interpreted as increased temporal consistency, not automatically as increased metric accuracy.

The adaptive schedule in Fig. 12 varies keyframe usage and optimization effort with scene conditions. This behavior is consistent with the design goal of allocating more computational resources when the scene is difficult. A future controlled ablation should isolate each controller action to determine which adaptations are responsible for changes in trajectory or reconstruction.

5.4 Qualitative Reconstruction

Figure 13 shows an underwater reconstruction example. Missing or dark silhouette regions are associated with occlusion, weak depth support, or insufficient viewpoints. These failure regions are important because they show where semantic supervision does not compensate for missing geometric information.

Table 3: Quantitative results explicitly supported by the current experiment files.

Dataset / protocol	Metric	Value
Replica Room0 ablation	PSNR, semantic mask	34.25 dB
Replica Room0 ablation	PSNR, no mask	20.18 dB
Replica Room0 ablation	Depth L1, mask	≈ 0.00
Replica Room0 ablation	Depth L1, no mask	0.23
Replica Room0 ablation	IoU, mask	0.8916
KUMP, 200 frames	ATE RMSE	41.79 cm
KUMP, 200 frames	RGB PSNR	19.12 dB
KUMP, 200 frames	Depth L1	7.79 cm

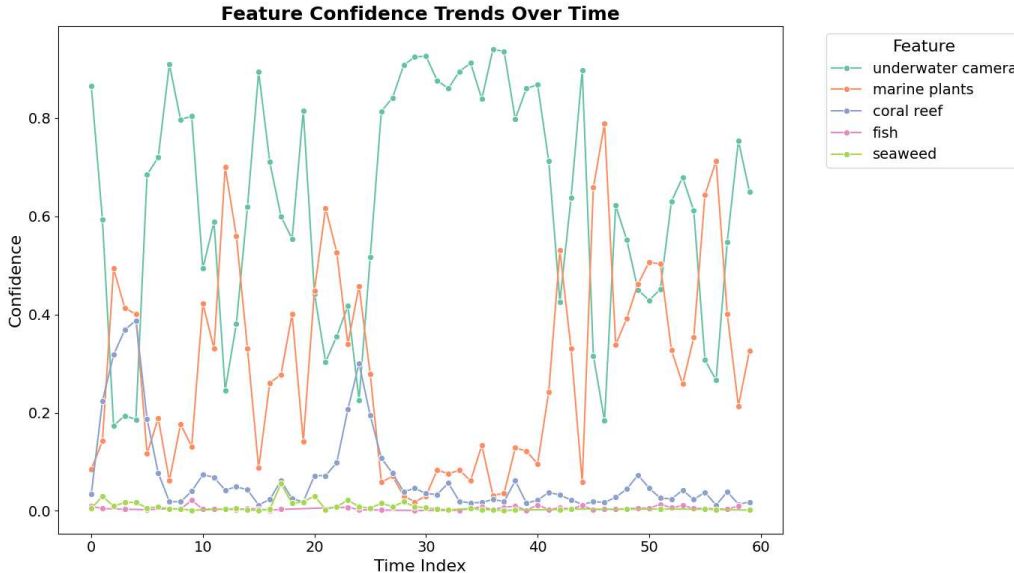


Figure 8: Temporal variation of semantic-prompt confidence in the underwater sequence.

6 Discussion

The experiments support three observations. First, semantic masking can substantially improve reconstruction in a controlled Replica example. Second, semantic prompts provide interpretable scene descriptors that can be connected to estimator configuration. Third, the complete system can operate on a real underwater sequence and produce a coherent trajectory and dense reconstruction despite visually difficult conditions.

The current evidence is nevertheless narrower than the original draft suggested. In particular, no matched ORB-SLAM3 or DROID-SLAM experiments are included in the provided results, so the manuscript should not claim 28% ATE or 22% RPE superiority over those systems. Similarly, an additional processing cost of roughly 0.23 s per frame, as observed in the current prototype, corresponds to only a few frames per second if incurred serially. It is therefore more accurate to describe the implementation as online or near-real-time on a laptop GPU rather than as high-rate real-time SLAM. Parallel execution, asynchronous semantic updates, and smaller models are natural routes to higher throughput.

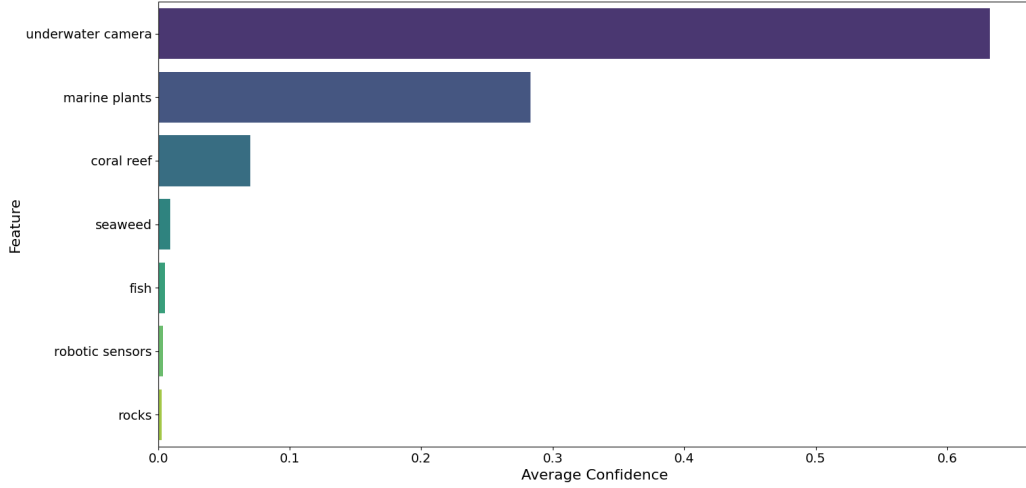


Figure 9: Most prominent semantic prompts in the underwater sequence.

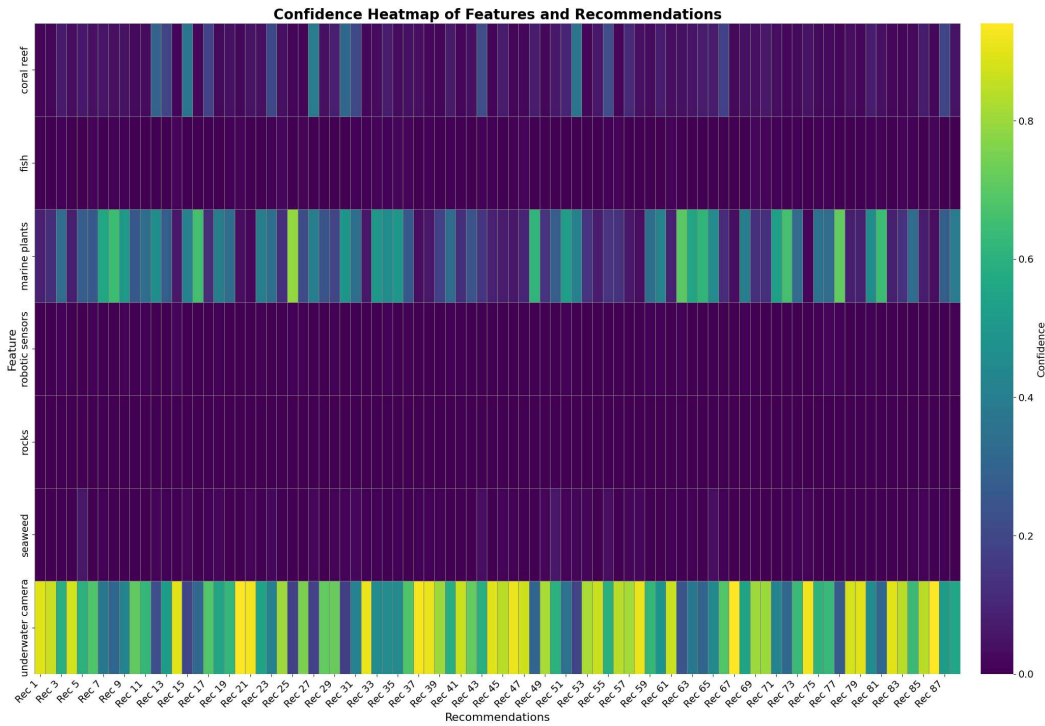


Figure 10: Relationship between semantic confidence and adaptive mapping recommendations.

A second important point concerns the role of the LLM. Free-form LLM control would make the system difficult to reproduce and could cause unstable parameter changes. The bounded action formulation in this revision makes the intended architecture clearer: the LLM is a high-level proposal mechanism, while deterministic parsing, limits, and geometric losses retain authority over the actual estimator.

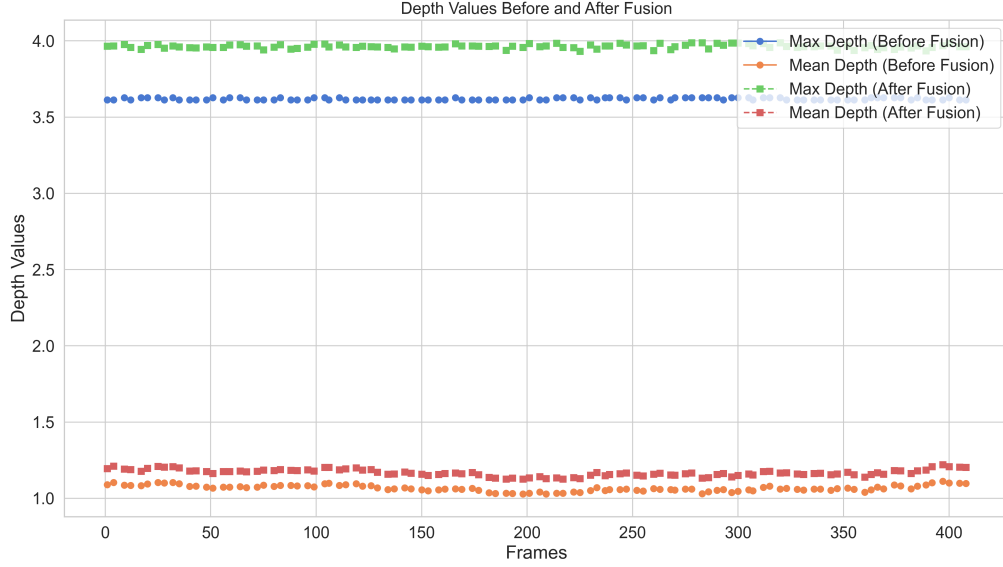


Figure 11: Depth statistics across frames before and after semantic confidence processing.

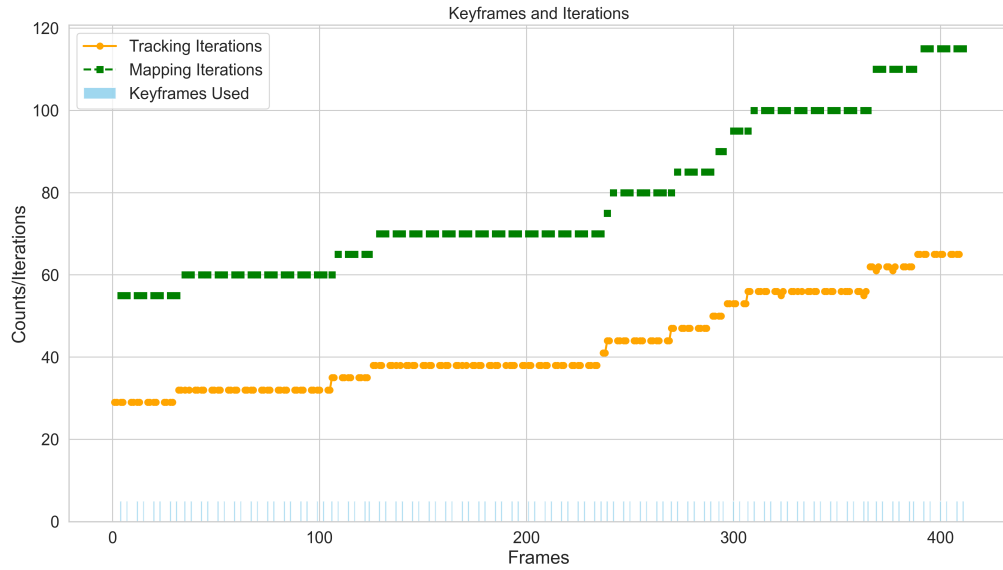


Figure 12: Evolution of keyframe usage and optimization iterations under adaptive control.

7 Limitations and Future Work

The present study has five main limitations. First, the quantitative evaluation is based on limited sequences; longer trajectories and more Replica scenes are needed to characterize drift and repeatability. Second, matched baseline runs against established geometric and learned SLAM systems are required before comparative performance claims can be made. Third, underwater semantic labels are obtained from a general-purpose vision-language model and may be poorly calibrated under strong domain shift. Fourth, monocular depth does not provide guaranteed metric scale, and additional scale alignment or sensor fusion should be incorporated for field deployment. Fifth, the current language-model and depth components add substantial computational overhead for

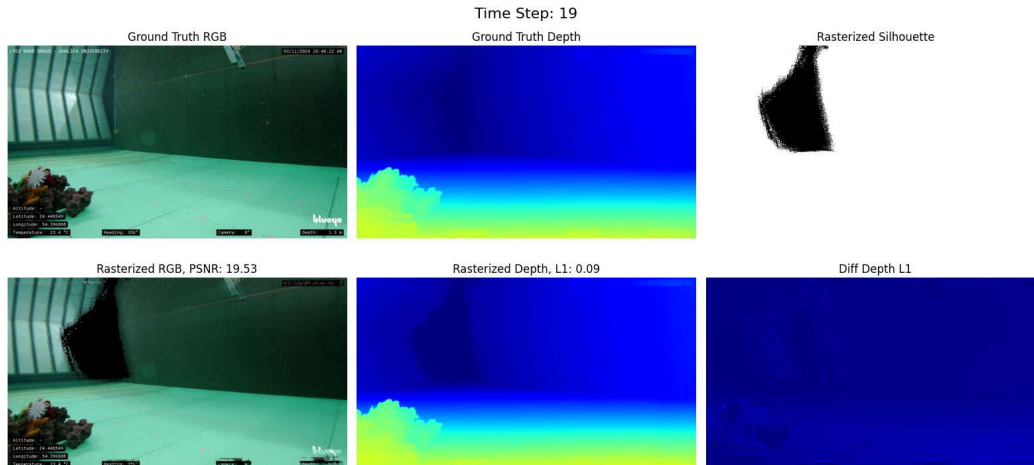


Figure 13: Representative underwater reconstruction output at time index 19.

embedded platforms.

Future work will therefore focus on (i) a multi-sequence ablation covering depth-only, semantics-only, fixed-parameter, and adaptive variants; (ii) matched ORB-SLAM3, DROID-SLAM, and contemporary neural/Gaussian mapping baselines; (iii) uncertainty calibration for underwater depth and semantics; (iv) asynchronous controller updates at a lower rate than geometric tracking; and (v) validation on longer underwater datasets and field missions with independent trajectory references.

8 Conclusion

This paper presented SAFARI, a semantic-depth adaptive SLAM framework for visually challenging robotic environments. The system combines foundation-model depth prediction, zero-shot visual semantics, a confidence-weighted depth objective, and bounded scene-conditioned adjustment of selected mapping parameters. The design keeps learned reasoning in a supervisory role while geometric optimization remains responsible for trajectory and map estimation. Experiments on Replica and an underwater ROV sequence show that semantic guidance can improve reconstruction in the evaluated indoor ablation and that context-conditioned adaptation is feasible during underwater mapping. The results also identify the work still required for a stronger comparative claim: broader sequences, matched SLAM baselines, calibrated uncertainty, and a faster implementation. These directions provide a concrete path from the current prototype toward robust adaptive SLAM for field robotics.

Data and Code Availability

The current manuscript package does not provide a verified public repository URL. Code and the KUMP sequence should be released with a stable archival link before final publication where licensing and institutional requirements permit. A blinded or private repository may be used during peer review if required by the journal.

Declarations

Competing interests. The authors should insert the appropriate competing-interest declaration at submission.

Funding. Funding information should be inserted if applicable.

Author contributions. A contributor statement should be added before final submission in accordance with Springer Nature requirements.

Acknowledgements

The authors may acknowledge institutional or technical support here. Any use of generative AI or language-editing tools should be disclosed according to the journal’s current editorial policy at the time of submission.

References

- [1] O. Alvarez-Tunon, H. Kanner, L.R. Marnet, and E. Kayacan. Mimir-uw: A multipurpose synthetic dataset for underwater navigation and inspection. In *IEEE International Conference on Intelligent Robots and Systems*, 2023.
- [2] M.G. Arenas, T. Xiao, S. Singh, and A. Zeng. How to prompt your robot: A promptbook for manipulation skills with code as policies. In *IEEE International Conference on Robotics and Automation, ICRA*, 2024.
- [3] Y. Cho and A. Kim. Visibility enhancement for underwater visual slam based on underwater light scattering model. In *Proceedings - IEEE International Conference on Robotics and Automation*, 2017.
- [4] W. Ding, Y. Wang, R. Zhang, and M. Wu. Tightly-coupled lidar-visual-inertial slam considering 3d-2d line correspondences. In *Proceedings of the 7th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, pages 121–130, 2023.
- [5] Ryan M Eustice, Hanumant Singh, and John J Leonard. Exactly sparse delayed-state filters for view-based slam. *IEEE Transactions on Robotics*, 22(6):1100–1114, 2006.
- [6] B. Fu and X. Yao. Visual slam system integrating feature extraction based on vision transformer and cloud-edge technology. In *International Conference on Intelligent Computing and Robotics (ICICR)*, 2024.
- [7] C. Guo, Y. Liu, Y. Luo, and Q. Zhang. Research progress in the application of image semantic information in visual slam. *Cehui Xuebao/Acta Geodaetica et Cartographica Sinica*, 2024.
- [8] S. Han, M.U. Ahmed, and P.K. Rhee. Monocular slam and obstacle removal for indoor navigation. In *International Conference on Machine Learning and Data Engineering*, 2018.
- [9] S. Han and Z. Xi. Dynamic scene semantics slam based on semantic segmentation. *IEEE Access*, 2020.
- [10] Nikhil Keetha, Jayesh Karhade, Krishna Murthy Jatavallabhula, Guanhua Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. *arXiv preprint arXiv:2312.02126*, 2023.

- [11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.
- [12] F. Li, X. Li, Y. Peng, and Y. Zhai. Maximum information transfer and minimum loss dehazing for underwater image restoration. *IEEE Journal of Oceanic Engineering*, 2024.
- [13] Ian Mahon, Stefan B Williams, Oscar Pizarro, and Matthew Johnson-Roberson. Efficient view-based slam using visual loop closures. *IEEE Transactions on Robotics*, 24(5):1002–1014, 2008.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.
- [15] S. Rahman, A. Quattrini Li, and I. Rekleitis. Svin2: A multi-sensor fusion-based underwater slam system. *International Journal of Robotics Research*, 2022.
- [16] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Stephen Green, Jakob J Engel, Raul Mur-Artal, Chen Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. In *arXiv preprint arXiv:1906.05797*, 2019.
- [17] A. Takmaz, D.P. Paudel, T. Probst, and L. Van Gool. Unsupervised monocular depth reconstruction of non-rigid scenes. In *International Conference on 3D Vision, 3DV*, 2021.
- [18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [19] Q. Ul Islam, H. Ibrahim, P.K. Chin, and F. Khozaei. Advancing real-world visual slam: Integrating adaptive segmentation with dynamic object detection for enhanced environmental perception. *Expert Systems with Applications*, 2024.
- [20] Z. Wen, Y. Zhao, F. Gao, and J. Dong. Nuam-net: A novel underwater image enhancement attention mechanism network. *Journal of Marine Science and Engineering*, 2024.
- [21] W. Wu, L. Guo, H. Gao, and Z. Chen. Yolo-slam: A semantic slam system towards dynamic environment with geometric constraint. *Neural Computing and Applications*, 2022.
- [22] Z. Xin, Z. Wang, Z. Yu, and B. Zheng. Ull-slam: underwater low-light enhancement for the front-end of visual slam. *Frontiers in Marine Science*, 2023.
- [23] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024.
- [24] M. Zhong, C. Hong, Z. Jia, and Z. Wang. Dynatm-slam: Fast filtering of dynamic feature points and object-based localization in dynamic indoor environments. *Robotics and Autonomous Systems*, 2024.