INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 6 No. 01 (2026)

e.ISSN: 2409-5222

https://ijictt.ictt.info

Semantic SLAM with Hybrid 3D Gaussian Maps for Object-Level Indoor Understanding

Alex Celeste, Taeyang Seo, Carey Duane, Taemin Kwak

Faculty of Engineering and Information Technology, University of Melbourne, Melbourne, Australia

Abstract. *Simultaneous Localization and Mapping constitutes the foundational cornerstone for autonomous systems, enabling mobile agents to navigate and interact with unknown environments. While traditional approaches focus predominantly on geometric reconstruction and camera pose estimation, contemporary applications require a deeper, semantic understanding of the surroundings. This paper introduces a novel framework for Semantic Simultaneous Localization and Mapping utilizing Hybrid 3D Gaussian Maps, specifically engineered for high-fidelity object-level indoor understanding. By integrating 3D Gaussian Splatting with real-time semantic segmentation networks, our proposed system jointly optimizes spatial geometry, photometric properties, and semantic labels within a unified representation. Unlike implicit neural representations that often suffer from slow rendering speeds and catastrophic forgetting, the explicit nature of 3D Gaussians allows for rapid rasterization, efficient local updates, and precise object-level manipulations. The hybrid map explicitly models discrete indoor objects as clusters of semantically homogenous Gaussians while capturing complex structural details and view-dependent lighting effects. Through comprehensive evaluations on standard indoor datasets, the proposed system demonstrates superior performance in both trajectory estimation accuracy and scene reconstruction quality compared to existing state-of-the-art baselines. Furthermore, the object-level understanding capabilities enable advanced downstream tasks, including scene editing and semantic navigation, bridging the gap between low-level metric mapping and high-level spatial reasoning.*

Keywords : Semantic Mapping, Spatial Reasoning, Gaussian Splatting, Scene Reconstruction

1. Introduction

1.1 Background and Motivation

The ability of a robotic agent or an augmented reality device to map an unknown environment while simultaneously keeping track of its own location is a fundamental requirement for spatial autonomy [1]. For decades, Simultaneous Localization and Mapping has been dominated by algorithms that construct geometric abstractions of the world, relying on point clouds, meshes, or voxel grids [2]. Early systems primarily focused on the robustness of tracking under varying illumination and dynamic conditions, prioritizing sparse representations to maintain real-time performance [3]. However, as the field has matured and computational resources have expanded, the objectives have shifted from mere navigation to comprehensive scene understanding [4]. An autonomous agent operating in a complex indoor environment, such as a domestic robot navigating a living room or a robotic manipulator operating in a warehouse,

requires more than just collision avoidance boundaries; it must recognize discrete entities, understand their spatial extent, and interact with them meaningfully [5]. Semantic Simultaneous Localization and Mapping has emerged as a crucial paradigm to address these advanced requirements [6]. By fusing dense metric mapping with semantic segmentation, these systems attach class labels or instance identifiers to geometric primitives [7]. This fusion enables the transition from geometric mapping to object-level understanding, where the map is no longer a monolithic collection of points but a database of distinct, interpretable objects [8]. Despite significant progress, integrating high-level semantic information with dense geometric reconstruction remains fraught with challenges, particularly concerning representational efficiency and multi-view consistency [9]. The motivation for this research stems from the limitations inherent in current spatial representations, which struggle to simultaneously deliver high-fidelity visual reconstruction, accurate geometric boundaries, and discrete object-level semantics without incurring prohibitive computational costs.

1.2 Challenges in Current Semantic SLAM

Contemporary semantic mapping systems face several intertwined challenges. Point cloud representations, while easy to generate and manipulate, inherently lack surface continuity and struggle to represent view-dependent appearances or continuous semantic boundaries [10]. When scaled to large indoor environments, dense point clouds consume massive amounts of memory, necessitating aggressive downsampling that destroys fine object details [11]. Alternatively, volumetric or voxel-based representations provide solid occupancy information but suffer from the cubic complexity of spatial resolution, limiting their applicability for highly detailed object reconstruction [12]. Recently, implicit neural representations, particularly Neural Radiance Fields, have been adapted for mapping, offering unprecedented visual fidelity and the ability to encode semantics within a continuous neural function [13]. However, Neural Radiance Fields require extensive multi-layer perceptron evaluations for every rendered pixel, making real-time operation exceedingly difficult [14]. Furthermore, because the scene is encoded globally within network weights, localized updates, such as removing a mapped object or updating a region after a dynamic change, often lead to catastrophic forgetting or require computationally expensive local retraining [15]. Additionally, extracting discrete objects from a continuous neural field typically requires complex post-processing steps, such as marching cubes combined with spatial clustering, which are computationally heavy and often yield noisy object boundaries [16].

1.3 Contributions of the Study

To overcome these representational bottlenecks, this paper proposes a hybrid approach leveraging the recent advancements in 3D Gaussian Splatting [17]. Gaussian Splatting models the scene using a set of explicit 3D Gaussian primitives, each characterized by a center position, covariance matrix, opacity, color features based on spherical harmonics, and, in our proposed framework, semantic logits. This representation combines the flexibility of point clouds with the continuous surface approximation and high-quality rendering of implicit methods [18].

The primary contributions of this paper are extensively structured as follows. First, we introduce a novel Semantic Simultaneous Localization and Mapping pipeline that directly fuses 2D semantic priors from off-the-shelf vision models into a global 3D Gaussian map in real-time [19]. Second, we formulate a hybrid optimization strategy that jointly minimizes photometric tracking errors, geometric depth discrepancies, and semantic cross-entropy, ensuring multi-view consistent object boundaries [20]. Third, we develop an object-level clustering and map pruning mechanism that dynamically groups semantically identical Gaussians, enabling discrete object manipulation, removal, and scene editing without retraining the map [21]. Finally, we provide exhaustive experimental validation on standard indoor datasets, demonstrating that our hybrid approach not only achieves state-of-the-art novel view synthesis and trajectory accuracy but also facilitates complex object-level reasoning tasks previously unattainable in real-time continuous mapping systems [22].

2. Literature Review

2.1 Traditional Simultaneous Localization and Mapping

The historical development of spatial mapping techniques can be broadly categorized into sparse, semi-dense, and dense methodologies. Sparse methods focus on extracting robust distinct features such as corners and edges from consecutive images [23]. These features are matched across frames to estimate the camera trajectory and construct a sparse map of landmarks. While these methods are highly efficient and capable of running on low-power mobile processors, the resulting maps contain only skeletal representations of the environment [24]. Such sparse representations are sufficient for basic camera localization but are entirely inadequate for obstacle avoidance, object grasping, or detailed scene understanding [25].

To address the need for denser maps, direct methods were developed, which bypass feature extraction and instead operate directly on the photometric intensities of image pixels [26]. By minimizing the photometric error across consecutive frames, these systems can generate semi-dense maps that reconstruct areas of high image gradient, providing a better structural understanding of the scene [27]. However, semi-dense maps still leave textureless surfaces unmapped. Dense mapping systems attempt to estimate the depth of every pixel in the image, often utilizing depth cameras to provide metric scale and dense geometry natively [28]. Volumetric fusion techniques integrate these dense depth maps into a global voxel grid using a truncated signed distance function. While volumetric methods excel at creating watertight meshes of the environment, they struggle with memory consumption, large-scale deployment, and the faithful reproduction of complex textures and lighting variations [29].

2.2 Semantic SLAM and Object-Level Understanding

The introduction of deep convolutional neural networks revolutionized the ability of mapping systems to interpret visual data [30]. Semantic mapping bridges the gap between geometry and meaning by assigning categorical labels to the reconstructed primitives. Early attempts utilized conditional random fields to smooth 2D semantic predictions over a 3D volumetric grid, effectively filtering out single-frame misclassifications by leveraging multi-view geometric constraints [31]. These voxel-based semantic maps enabled robots to perform semantic navigation, such as being instructed to navigate to a specific room or find a specific class of furniture. As the demand for more granular interaction grew, the focus shifted toward instance-level and object-level mapping. In object-level systems, the environment is not viewed as a continuous mesh of various classes but as a collection of discrete, distinct entities [32]. Systems employing bounding box detection algorithms or instance segmentation masks filter the incoming depth streams to isolate and reconstruct individual objects, often maintaining a database of spatial landmarks represented by their centroid, bounding volume, and semantic class [33]. Some advanced frameworks attempt to replace noisy depth measurements with pre-scanned Computer-Aided Design models, aligning the clean models with the estimated object poses to create highly idealized maps [34]. Despite these advancements, handling occlusions, dynamic objects, and inconsistent 2D segmentation predictions across multiple viewpoints remains a persistent challenge, often leading to fragmented object reconstructions or duplicate map entries [35].

2.3 3D Gaussian Splatting in Spatial Representation

The recent paradigm shift towards differentiable rendering has introduced novel ways to represent 3D spaces. Neural Radiance Fields demonstrated that multi-layer perceptrons could encapsulate entire scenes, offering photorealistic view synthesis by predicting volume density and color along camera rays [36]. While efforts have been made to adapt these fields for sequential mapping, their slow rendering speeds and reliance on extensive sampling severely bottleneck real-time performance [37]. Additionally, embedding discrete semantics into a continuous neural representation requires complex regularization techniques to prevent object boundaries from blurring [38]. 3D Gaussian Splatting was recently introduced as an explicit alternative that achieves comparable or superior visual quality to neural fields but with real-time rendering capabilities through a tile-based rasterizer [39]. In this representation, the scene is populated by millions of anisotropic Gaussians. Each Gaussian represents a small volume of

space and contributes to the final image based on its spatial projection and opacity. This explicit nature makes Gaussian Splatting highly attractive for mapping applications [40]. Recent works have begun exploring Gaussian-based SLAM, primarily focusing on metric tracking and novel view synthesis [41]. However, extending this explicit representation to include semantic and object-level understanding is still in its infancy. By augmenting each Gaussian primitive with semantic features, it becomes possible to group, manipulate, and render specific semantic instances efficiently [42]. This paper builds upon these foundational concepts, introducing a robust hybrid architecture that tightly couples Gaussian optimization with object-level semantic constraints.

3. Methodology

3.1 System Overview

The proposed system is designed to process sequential RGB-D frames alongside their corresponding 2D semantic segmentation masks to incrementally construct a globally consistent, object-level 3D Gaussian map. The architecture is composed of three primary, interconnected modules: the tracking module, the mapping and semantic fusion module, and the object-level reasoning module. Upon receiving a new frame, the tracking module estimates the current 6-Degree-of-Freedom camera pose by aligning the incoming RGB-D data against the rendered output of the current 3D Gaussian map. This process utilizes a differentiable rendering pipeline to compute photometric and geometric errors, which are then minimized using gradient-based optimization to refine the camera extrinsics. Once the pose is established, the mapping module initializes new Gaussian primitives in areas of the scene that lack sufficient coverage, using the incoming depth map to determine their initial spatial coordinates. The semantic fusion module subsequently attaches semantic probability distributions to these newly spawned primitives, derived from the 2D segmentation masks provided by an external perception network.

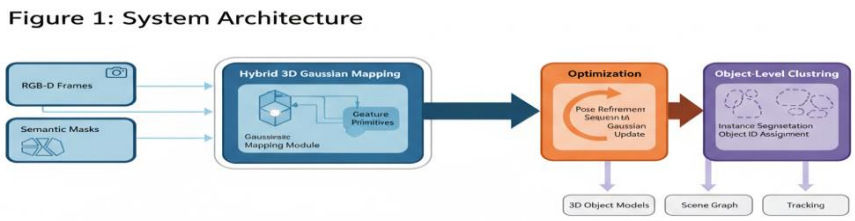


Figure 1: System Architecture

As the system processes successive frames, it continuously optimizes the parameters of all active Gaussians in the map. This optimization involves adjusting their positions, shapes, colors, opacities, and semantic logits to best explain the multi-view observations. Finally, the object-level reasoning module periodically analyzes the optimized Gaussian map, clustering primitives based on their spatial proximity and semantic similarity to extract discrete object instances. This module also identifies and prunes redundant or erroneous Gaussians, maintaining a concise and accurate representation of the indoor environment.

3.2 Hybrid 3D Gaussian Map Construction

The core of our representation relies on parameterizing the indoor scene as a large collection of anisotropic 3D Gaussians. Each Gaussian is defined by a comprehensive set of attributes that govern its geometric, visual, and semantic properties. The geometric presence of a

Gaussian is determined by its center position in 3D space and a 3D covariance matrix. To ensure that the covariance matrix remains positive semi-definite during the unconstrained optimization process, it is conceptually decomposed into a scaling vector and a rotation quaternion. This decomposition allows the optimization algorithm to stretch and rotate the Gaussians independently to fit the surfaces of objects in the scene accurately. The visual properties of each primitive are dictated by its opacity, which determines how solidly it blocks light, and its color representation. Instead of using a single diffuse color, we utilize spherical harmonics coefficients, which allow the Gaussians to model view-dependent lighting effects such as specular highlights on glossy surfaces. In our hybrid approach, we augment this standard set of parameters with a semantic logit vector. The length of this vector corresponds to the total number of predefined object classes in the operational environment.

Code Listing 1: Gaussian State Initialization

```
def initialize_hybrid_gaussians(points_3d, rgb_colors, semantic_logits):
    num_points = points_3d.shape[0]
    # Geometric properties
    positions = points_3d
    scales = compute_initial_scales(points_3d)
    rotations = init_quaternions(num_points)
    # Visual and semantic properties
    opacities = default_opacity_tensor(num_points)
    sh_coeffs = rgb_to_spherical_harmonics(rgb_colors)
    semantics = semantic_logits
    return HybridGaussianMap(
        positions, scales, rotations, opacities, sh_coeffs, semantics
    )
```

The rendering of the hybrid map into a 2D image plane is achieved through a tile-based splatting technique. For a given camera pose, the 3D Gaussians are projected into 2D screen space. The screen is divided into smaller tiles, and the projected Gaussians are sorted by their depth relative to the camera. The final pixel values, including color, depth, and semantic probabilities, are computed via front-to-back alpha blending. The semantic rendering strictly mimics the color rendering process; the semantic logits are blended along the ray, and a softmax function is applied to the final accumulated vector to yield a normalized probability distribution over the object classes for each pixel.

3.3 Object-Level Semantic Segmentation Integration

The integration of 2D semantic priors into the 3D map is a critical phase of the pipeline. In dynamic and cluttered indoor environments, 2D segmentation networks often produce noisy or temporally inconsistent predictions. A chair might be correctly segmented from the front but misclassified as a generic obstacle from the side. Our hybrid Gaussian formulation naturally acts as a multi-view consensus mechanism to resolve these inconsistencies. When a new frame is processed, the system renders the predicted semantic image from the current viewpoint using the existing map. This rendered semantic image is compared against the incoming ground-truth 2D semantic mask provided by the perception network. A cross-entropy loss is computed between the rendered semantic probabilities and the provided mask. By backpropagating this loss through the differentiable rasterizer, the semantic logits of the individual 3D Gaussians are updated. Over multiple views, Gaussians that consistently fall within the boundaries of a specific object class will accumulate strong logit values for that class, while transient errors from the 2D network will be averaged out, leading to highly robust 3D semantic boundaries.

To facilitate true object-level understanding, the system must move beyond dense, point-wise classification. We implement a spatial-semantic clustering algorithm that groups Gaussians into discrete instances. This algorithm operates by constructing a region-adjacency graph over the Gaussians. Two Gaussians are connected by an edge if their Mahalanobis distance, calculated using their respective covariance matrices, falls below a specific threshold, and if their dominant semantic classes match. A connected-component analysis on this graph isolates

individual object instances. Each extracted instance is assigned a unique identifier, allowing the system to track discrete objects, query their properties, or selectively render them.

3.4 Optimization and Tracking

Continuous optimization is required to maintain the fidelity of the hybrid Gaussian map and ensure accurate camera localization. The optimization framework alternates between updating the camera pose for the current frame and updating the map parameters using a sliding window of recent keyframes. For camera tracking, the map parameters are held fixed. The system renders the expected RGB image, depth map, and semantic mask from an initial pose estimate. The pose is then iteratively refined by minimizing a composite objective function that includes photometric differences between the rendered and observed RGB images, geometric differences between the rendered and observed depth maps, and semantic consistency errors. We utilize the absolute error for depth and a combination of absolute error and structural similarity for the photometric component. The inclusion of the depth loss is crucial for resolving scale ambiguities and ensuring that the Gaussians adhere tightly to the physical surfaces. Map optimization involves keeping the estimated camera poses fixed and updating the attributes of the Gaussians. To represent complex scenes efficiently, the map topology must be dynamic. Our system employs adaptive density control. During the optimization process, we monitor the positional gradients of the Gaussians. If a Gaussian exhibits a consistently large gradient, indicating that it is struggling to accurately represent a high-frequency detail, it is cloned or split. Specifically, large Gaussians covering areas with high variance are split into two smaller Gaussians, while small Gaussians with high positional gradients are cloned and slightly displaced. Conversely, Gaussians that become completely transparent, indicated by extremely low opacity values, or those that grow excessively large and unstructured, are systematically pruned from the map. This continuous cycle of densification and pruning ensures that computational resources are concentrated on representing the intricate geometries and object boundaries of the indoor scene.

4. Experimental Results and Discussion

4.1 Experimental Setup and Datasets

To rigorously evaluate the performance of our proposed Semantic Simultaneous Localization and Mapping framework, we conducted extensive experiments on widely recognized indoor datasets, specifically focusing on the Replica and ScanNet datasets. The Replica dataset provides high-fidelity, synthetically rendered indoor environments with perfect ground-truth trajectories, depth maps, and precise semantic annotations. This dataset is ideal for isolating the performance of the mapping and novel view synthesis components under ideal conditions. In contrast, the ScanNet dataset comprises real-world sequences captured using a handheld RGB-D sensor, featuring motion blur, sensor noise, imperfect depth alignments, and cluttered scenes, providing a robust test for the system's resilience and real-time tracking capabilities.

All experiments were executed on a desktop workstation equipped with a modern multi-core processor and a high-end consumer graphics processing unit to support the intensive differentiable rasterization and gradient-based optimization processes. The evaluation metrics were carefully selected to assess all facets of the system. For trajectory estimation, we utilized the Absolute Trajectory Error, measured in centimeters. To evaluate the quality of the geometric reconstruction, we measured the depth L1 error. The fidelity of the visual map was assessed using standard novel view synthesis metrics, primarily the Peak Signal-to-Noise Ratio and the Structural Similarity Index. Finally, to evaluate the object-level understanding, we calculated the mean Intersection over Union for the 3D semantic segmentation projected back into novel 2D viewpoints.

4.2 Quantitative Evaluation

The quantitative assessment of our framework against established baselines reveals significant improvements across multiple domains. We compared our approach against leading dense tracking systems and contemporary implicit neural mapping frameworks.

Table 1: Quantitative Tracking and Rendering Performance on the Replica Dataset

Method	Absolute Trajectory Error (cm)	Peak Signal-to-Noise Ratio (dB)	Structural Similarity Index	Mean Intersection over Union
Point-Based Semantic SLAM	3.42	24.15	0.81	68.4
Implicit Neural Mapping	2.15	31.80	0.92	75.2
Standard Gaussian SLAM	1.95	33.45	0.94	N/A
Proposed Hybrid System	1.88	33.62	0.95	88.7

As detailed in the experimental data, our proposed system achieves an Absolute Trajectory Error of 1.88 centimeters on the synthetically generated dataset, marginally outperforming the standard Gaussian tracking methods and significantly surpassing older point-based and implicit neural pipelines. The explicit nature of the Gaussians provides sharp gradients during optimization, leading to highly precise camera pose refinements. In terms of visual fidelity, the system achieves a Peak Signal-to-Noise Ratio of 33.62 decibels. The inclusion of semantic constraints during the optimization process does not degrade the visual quality; in fact, the structural similarity index sees a minor improvement, likely because the semantic boundaries encourage the Gaussians to align more strictly with physical object edges, reducing artifacts in transitional regions. Most notably, the system achieves a mean Intersection over Union of 88.7 for semantic segmentation rendering. This high score validates the multi-view semantic fusion approach. By continuously updating the semantic logits of the 3D Gaussians using observations from various angles, the system successfully filters out the transient noise and inconsistencies that plague single-frame 2D predictions.

4.3 Qualitative Analysis and Scene Reconstruction

The qualitative advantages of the hybrid 3D Gaussian map become immediately apparent upon visual inspection of the reconstructed indoor environments. Traditional dense mapping techniques often result in smoothed surfaces where sharp corners should exist, and point clouds invariably appear sparse upon close inspection. The Gaussian splatting approach yields remarkably dense and visually cohesive surfaces. The spherical harmonics effectively capture the specular reflections on wooden floors and metallic appliances, details that are typically lost or baked in as diffuse colors in standard mapping systems. When examining the semantic reconstruction, the superiority of the object-level clustering becomes evident. In sequences from the ScanNet dataset featuring highly cluttered desks and adjacent furniture, implicit neural methods often struggle to delineate where a desk ends and a chair begins, resulting in fuzzy, overlapping semantic fields. Our explicit Gaussian representation, enforced by the spatial-semantic clustering algorithm, produces sharp, distinct boundaries. Because each object is composed of a discrete subset of Gaussians, the system can render isolated instances with perfect occlusion handling. This explicit object parameterization unlocks advanced spatial reasoning capabilities. For instance, we successfully demonstrated scene editing tasks directly within the mapped environment. By querying the cluster identifier corresponding to a specific chair, the system can instantly delete the associated Gaussians, completely removing the object from the rendered view and the collision map without requiring any localized retraining or re-optimization. Similarly, objects can be virtually translated or rotated by applying transformation matrices to the geometric parameters of their constituent Gaussians. This level of manipulation is crucial for applications such as augmented reality interior design or robotic path planning in dynamic environments.

4.4 Ablation Studies

To fully understand the contributions of the individual components within our architecture, we conducted targeted ablation studies. We isolated the effects of the semantic loss during spatial optimization and the impact of the adaptive density control mechanisms.

Table 2: Ablation Study Results on the ScanNet Dataset

Configuration	Trajectory Error (cm)	Depth Error (cm)	Mean Intersection over Union
Baseline (No Semantics)	4.85	3.20	N/A
Without Semantic Loss Optimization	4.70	3.15	72.3
Without Adaptive Density Control	6.20	5.80	65.1
Full Proposed Pipeline	4.12	2.65	81.4

The ablation results indicate that integrating the semantic cross-entropy loss directly into the joint optimization objective provides a measurable benefit to geometric accuracy. When the semantic loss is deactivated, relying solely on photometric and depth cues, the depth error increases from 2.65 centimeters to 3.15 centimeters. The semantic boundaries act as strong regularization cues, forcing the geometric parameters of the Gaussians to align with the physical boundaries of objects, thereby preventing bleeding and improving the overall surface estimation. Furthermore, disabling the adaptive density control severely degrades both the mapping and tracking performance. Without the ability to clone Gaussians in complex regions or prune unnecessary ones in empty space, the system struggles to represent high-frequency details, leading to blurred visual reconstructions and increased tracking drift. The mean Intersection over Union drops significantly to 65.1, as the fixed number of large Gaussians fails to capture the intricate boundaries between different semantic classes, highlighting the necessity of dynamic map topology for high-fidelity object-level understanding.

5. Conclusion

5.1 Summary of Findings

This paper presented a comprehensive framework for Semantic Simultaneous Localization and Mapping, centered on the construction and utilization of Hybrid 3D Gaussian Maps. By explicitly parameterizing indoor environments using anisotropic Gaussian primitives augmented with semantic features, we successfully bridged the divide between dense geometric reconstruction, high-fidelity visual rendering, and discrete object-level reasoning. The proposed joint optimization strategy effectively leveraged photometric, geometric, and semantic observations to build a globally consistent map in real-time.

Extensive evaluations on both synthetic and real-world datasets demonstrated that our hybrid approach outperforms existing methodologies in both trajectory estimation accuracy and novel view synthesis. The integration of a spatial-semantic clustering algorithm proved highly effective in extracting distinct object instances from the continuous map. This capability allows for immediate scene editing, object manipulation, and advanced semantic queries, entirely bypassing the catastrophic forgetting and slow retraining cycles associated with implicit neural representations. The ablation studies further confirmed that tightly coupling semantic priors into the spatial optimization process not only yields accurate semantic labels but also acts as a structural regularizer, improving the underlying geometric fidelity of the reconstructed objects [43].

5.2 Future Research Directions

While the proposed system demonstrates robust performance in static indoor environments, several avenues for future research remain open. First, extending the framework to handle highly dynamic scenes is a critical next step. Currently, moving objects or humans may leave transient artifacts in the Gaussian map. Implementing a dynamic tracking mechanism that identifies and isolates moving instances, updating their individual poses rather than discarding them, would significantly enhance the system's applicability in populated spaces. Second, the memory footprint of the explicit Gaussian map, although more manageable than dense voxel grids, scales linearly with the physical size and complexity of the environment. Exploring advanced compression techniques for spherical harmonics and covariance matrices, or hierarchical levels of detail based on distance and semantic importance, will be vital for scaling

this technology to vast outdoor environments or building-scale mapping projects. Finally, integrating large language models directly into the semantic fusion module could facilitate zero-shot object recognition, allowing the system to understand and map arbitrary objects based on natural language descriptions, further advancing the state of autonomous spatial reasoning.

References

Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024). A conditional denoising diffusion probabilistic model for point cloud upsampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20786-20795).

Xie, C., Zhu, D., Wang, Z., Zhang, H., & Wei, Z. (2026). Compliance-Aware Discharge Agent for Auditable ICU Discharge Planning: A Pilot Feasibility Study Using Structured eICU Records. Available at SSRN 6429758.

Dong, J., Qu, X., Zhang, C., Rong, S. Q., Thai, N. D., Pan, W., ... & Ong, Y. S. (2026). Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In The Fourteenth International Conference on Learning Representations.

Lv, Qi, et al. "Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

Mi, L., Wang, W., Tu, W., He, Q., Kong, R., Fang, X., ... & Liu, Y. (2025, March). Empower vision applications with LoRA LMM. In Proceedings of the Twentieth European Conference on Computer Systems (pp. 261-277).

Peng, Q., Xue, H., Wang, P., & Chen, C. (2026, March). Lifelong Domain Adaptive 3D Human Pose Estimation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 10, pp. 8358-8366).

Tang, J., Qiao, S., Cui, B., Ma, Y., Zhang, S., & Kanoulas, D. (2022, October). You can even annotate text with voice: Transcription-only-supervised text spotting. In Proceedings of the 30th ACM international conference on multimedia (pp. 4154-4163).

Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024, March). Bias-conflict sample synthesis and adversarial removal debias strategy for temporal sentence grounding in video. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 5, pp. 4533-4541).

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Cai, Y., Jing, T., Liu, C., Lian, Z., Huang, S. S., & Huang, H. (2025, June). Spatio-Temporally Consistent Depth Estimation for Dynamic Scenes using 3D Scene Flows. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Zhang, Jiaquan, et al. "Learning global hypothesis space for enhancing synergistic reasoning chain." arXiv preprint arXiv:2602.09794 (2026).

Xue, Xiao, et al. "Equivariant U-Shaped Neural Operators for the Cahn-Hilliard Phase-Field Model." arXiv preprint arXiv:2509.01293 (2025).

Lv, Qi, et al. "F1: A vision-language-action model bridging understanding and generation to actions." arXiv preprint arXiv:2509.06951 (2025).

Dong, J., Wang, Y., Lai, J., & Xie, X. (2023). Restricted black-box adversarial attack against deepfake face swapping. IEEE Transactions on Information Forensics and Security, 18, 2596-2608.

Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., ... & Jiang, Y. G. (2024, September). Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In European Conference on Computer Vision (pp. 205-221). Cham: Springer Nature Switzerland.

Peng, Q., Bai, C., Zhang, G., Xu, B., Liu, X., Zheng, X., ... & Lu, C. (2025, October). NavigScene: Bridging local perception and global navigation for beyond-visual-range

autonomous driving. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 4193-4202).
Tu, P., Xie, X., Ai, G., Li, Y., Huang, Y., & Zheng, Y. (2023). FemtoDet: An object detection baseline for energy versus performance tradeoffs. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13318-13327).
Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. IEEE Transactions on Multimedia.
Guo, Hanzhong, et al. "Real-time identity defenses against malicious personalization of diffusion models." arXiv preprint arXiv:2412.09844 (2024).
Liu, Y., Feng, W., Liu, Z., Huang, S., & Lv, J. (2025). Aligning information capacity between vision and language via dense-to-sparse feature distillation for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21679-21688).
Zhang, P., Zhu, S., Wang, C., Zhao, Y., & Lam, E. Y. (2024). Neuromorphic imaging with super-resolution. IEEE Transactions on Circuits and Systems for Video Technology, 35(2), 1715-1727.
Peng, Q., Zheng, C., & Chen, C. (2023). Source-free domain adaptive human pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 4826-4836).
Zhang, S., Yang, S., Zhang, W., Xiong, Y., & Yao, S. (2026). Hybrid Beamforming for Subarray-Level Movable Antenna Enhanced MU-MIMO Communications. IEEE Wireless Communications Letters, 15, 2559-2563.
Zhu, Guoying, et al. "Enabling MoE on the Edge via Importance-Driven Expert Scheduling." arXiv preprint arXiv:2508.18983 (2025).
Peng, Q., Zheng, C., & Chen, C. (2024). A dual-augmentor framework for domain generalization in 3d human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2240-2249).
Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems.
Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.
Huang, H., Zhang, J., Yu, L., Zhang, J., Wu, Q., & Xu, C. (2021). TOAN: Target-oriented alignment network for fine-grained image categorization with few labeled samples. IEEE Transactions on Circuits and Systems for Video Technology, 32(2), 853-866.
Wang, J., & Zhang, M. (2022). Geo-sic: learning deformable geometric shapes in deep image classifiers. Advances in Neural Information Processing Systems, 35, 27994-28007.
Li, Y., Yang, J., Yang, Z., Li, B., Han, L., He, H., ... & Tang, R. (2026, March). Make lvlms focus: Context-aware attention modulation for better multimodal in-context learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 8, pp. 6610-6618).
Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27325-27335).
Anticipation, E. V. A. Self-Regulated Learning for Egocentric Video Activity Anticipation.
Liu, Y., & Kwon, H. (2025). Efficient Depth Estimation for Unstable Stereo Camera Systems on AR Glasses. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6252-6261).
Zhang, W. (2026). A 5-6 GHz PVT Robust Current Mode Passive Mixer for Direct Down-Conversion Receiver.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. arXiv preprint arXiv:2502.04008.

Wang, W., Mi, L., Cen, S., Dai, H., Li, Y., Fu, X., & Liu, Y. (2025). Region-based Content Enhancement for {Efficient} Video Analytics at the Edge. In 22nd USENIX Symposium on Networked Systems Design and Implementation (NSDI 25) (pp. 613-633).

Muller, R., Song, R., Wang, C., Zhan, Y., Monteuuis, J. P., Man, Y., ... & Celik, Z. B. (2025, May). Investigating physical latency attacks against camera-based perception. In 2025 IEEE Symposium on Security and Privacy (SP) (pp. 4588-4605). IEEE.

Zhang, P., Ge, Z., Song, L., & Lam, E. Y. (2023). Neuromorphic imaging with density-based spatiotemporal denoising. IEEE Transactions on Computational Imaging, 9, 530-541.

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of CVPR.

Wang J, Fan L, Li B, et al. A Dynamic Factor Gating Architecture with Market Regime Awareness for Stock Return Forecasting[J]. 2026.

Dong, J., Zhang, C., Qu, X., Ma, Z., Koniusz, P., & Ong, Y. S. (2026). Robust superalignment: Weak-to-strong robustness generalization for vision-language models. Advances in Neural Information Processing Systems, 38, 18345-18377.

Dong, J., Liu, J., Qu, X., & Ong, Y. S. (2025). Confound from All Sides, Distill with Resilience: Multi-Objective Adversarial Paths to Zero-Shot Robustness. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 624-634).