ISSN: 2519-9439

Sparse Point-Cloud Semantic Understanding for Autonomous Mobile Sensing with Cross-Sensor Scene-Graph Reasoning in Complex Environments

Yuxuan Ma; Kit Ng

School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

Abstract

The rapid advancement of autonomous mobile sensing heavily relies on the ability of robotic systems to perceive and comprehend complex, dynamic environments. Among the various sensing modalities, Light Detection and Ranging provides critical three-dimensional structural information, yet it inherently suffers from spatial sparsity, particularly at extended ranges. This paper introduces a comprehensive framework for sparse point-cloud semantic understanding, augmented by cross-sensor scene-graph reasoning, designed specifically for autonomous mobile platforms operating in intricate real-world conditions. By integrating high-resolution visual data from cameras with sparse geometric data from laser scanners, the proposed methodology bridges the semantic gap that traditionally limits unimodal perception systems. We propose a novel architecture that extracts fine-grained local geometric features from sparse point clouds and fuses them with dense visual embeddings using a multi-modal attention mechanism. Furthermore, we construct a dynamic three-dimensional scene graph where nodes represent semantic entities and edges encapsulate their spatiotemporal relationships. This graph-based reasoning allows the system to infer complex interactions, such as occlusion dependencies and trajectory forecasting, which are crucial for safe navigation. Extensive evaluations on standard autonomous driving datasets demonstrate the superior performance of our approach in semantic segmentation and relational reasoning tasks compared to existing state-of-the-art methods. The integration of structural topology through scene graphs significantly enhances the robustness of semantic understanding in adverse weather and low-lighting conditions, paving the way for more resilient autonomous mobile sensing systems.

Keywords

Sparse Point Clouds, Scene Graph Reasoning, Cross-Sensor Fusion, Autonomous Mobile Sensing, Semantic Understanding.

1.Introduction

1.1 Background on Autonomous Mobile Sensing

The deployment of autonomous mobile systems, including self-driving vehicles, uncrewed aerial vehicles, and terrestrial delivery robots, requires a profound capacity to perceive, interpret, and navigate highly complex environments. Reliable perception forms the foundational pillar of any autonomous system, dictating the efficacy of subsequent decision-

ISSN: 2519-9439

making and trajectory planning algorithms. In the contemporary landscape of mobile sensing, platforms are typically equipped with a diverse suite of heterogeneous sensors designed to capture complementary aspects of the surrounding operational domain [36]. Light Detection and Ranging sensors have emerged as the primary modality for acquiring high-fidelity spatial configurations, offering direct distance measurements that are theoretically invariant to ambient illumination fluctuations [1]. Simultaneously, high-definition optical cameras provide dense, rich textural and color information, which is indispensable for recognizing fine-grained semantic details such as traffic sign symbols, lane markings, and pedestrian intent [2]. The overarching goal of autonomous mobile sensing is to synthesize these disparate data streams into a cohesive, actionable representation of the environment [3]. However, as platforms are deployed in increasingly unstructured and chaotic urban settings, the raw aggregation of sensory data is no longer sufficient. It becomes imperative to develop sophisticated perceptual models capable of moving beyond mere object detection toward holistic semantic understanding and relational reasoning [4].

1.2 Challenges in Sparse Point-Cloud Understanding

Despite the undeniable geometric advantages provided by laser scanning technologies, the processing of three-dimensional point clouds presents a unique set of computational and conceptual challenges. The most prominent of these is spatial sparsity. Unlike image data, which is arranged in a dense, regular grid of pixels, point clouds are fundamentally unstructured and irregularly distributed across the continuous three-dimensional space [5]. The density of captured points decreases quadratically as a function of the distance from the sensor, resulting in highly sparse representations for distant objects. For instance, a pedestrian located fifty meters away from an autonomous vehicle may be represented by fewer than ten discrete points, rendering traditional geometric feature extraction methods highly unreliable [6]. This sparsity is further exacerbated by the physical properties of the scanned surfaces, where reflective, transparent, or highly absorptive materials can cause erratic beam returns, leading to missing structural data or ghosting artifacts [7]. Traditional deep learning architectures, such as standard convolutional neural networks, are ill-suited for directly processing this irregular data format. Early attempts to mitigate this involved transforming the continuous point cloud into discrete intermediate representations, such as three-dimensional voxel grids or two-dimensional range images [8]. While voxelization imposes a regular grid structure that facilitates the application of three-dimensional convolutions, it introduces a severe quantization error and results in immense computational overhead due to the cubic growth of memory requirements relative to spatial resolution [9]. Conversely, projecting point clouds into two-dimensional panoramic views sacrifices critical geometric depth relationships, often leading to performance degradation in heavily occluded scenes [10]. Consequently, advancing the semantic understanding of sparse point clouds requires native architectures capable of extracting robust hierarchical features directly from the raw, unstructured point sets while maintaining computational efficiency and resilience to extreme variations in point density.

1.3 Objectives and Contributions

To address the limitations of unimodal sparse point cloud processing, this paper explores the paradigm of cross-sensor scene-graph reasoning. A scene graph is a structured data representation that encodes an environment not merely as a collection of isolated objects, but as a complex network of relational dependencies [11]. In this graphical topology, nodes correspond to distinct semantic entities, such as vehicles, pedestrians, and infrastructure elements, while the interconnecting edges represent spatial, semantic, or temporal relationships, such as a pedestrian crossing in front of a vehicle [12]. Transitioning from object-level detection to relational-level reasoning is paramount for complex autonomous navigation,

ISSN: 2519-9439

as it allows the system to predict contextual behaviors and infer the presence of occluded entities based on surrounding environmental cues [13]. The primary objective of this research is to formulate a comprehensive methodology that leverages the complementary strengths of sparse geometric data and dense visual data to construct dynamic, three-dimensional scene graphs [14]. By combining cross-sensor feature fusion with advanced graph neural networks, we aim to significantly elevate the semantic understanding capabilities of autonomous mobile platforms [15]. The core contributions delineated in this paper are manifold. First, we propose an adaptive spatial encoding module designed to extract highly discriminative geometric features from extremely sparse point cloud regions without relying on computationally prohibitive dense voxelization. Second, we introduce a novel cross-sensor multi-modal attention framework that effectively resolves the spatial misalignment between three-dimensional lidar coordinates and two-dimensional camera pixels, enabling seamless feature aggregation at the object level. Third, we formalize a deep relational reasoning architecture that dynamically generates comprehensive three-dimensional scene graphs, providing a rich, context-aware representation of complex operational environments. Finally, we provide an extensive empirical analysis demonstrating the efficacy and computational feasibility of the proposed system under challenging real-world scenarios.

2. Literature Review

2.1 Point Cloud Semantic Segmentation

The domain of three-dimensional semantic segmentation has witnessed immense paradigm shifts, largely driven by the advent of deep learning architectures tailored for unstructured geometric data. Early methodologies predominantly relied on handcrafted feature descriptors, such as Fast Point Feature Histograms, which were subsequently fed into traditional machine learning classifiers like Support Vector Machines or Random Forests [16]. While computationally lightweight, these approaches struggled to capture the complex, hierarchical abstractions necessary for robust semantic understanding in cluttered urban environments [17]. The introduction of deep neural networks capable of directly consuming raw point sets revolutionized the field. Architectures that utilize shared multi-layer perceptrons and symmetric aggregation functions demonstrated that it is possible to achieve permutation invariance while extracting rich global and local features [18]. However, these point-based methods frequently encounter scalability bottlenecks when processing the massive point clouds generated by modern autonomous platforms, which can exceed hundreds of thousands of points per frame. To mitigate these scalability issues, researchers explored sparse convolutional networks [19]. By constraining convolution operations to only the occupied regions of a voxelized space, sparse convolutions drastically reduce both memory consumption and computational latency compared to dense volumetric approaches [20]. Despite these advancements, handling extreme sparsity at long distances remains a critical vulnerability. When structural cues are insufficient, purely geometric networks often misclassify entities due to the absence of distinguishing morphological features [21]. This inherent limitation of unimodal point cloud processing forms the primary motivation for exploring multi-modal fusion strategies that incorporate rich visual textures to supplement missing geometric information.

2.2 Scene Graph Generation in 3D Environments

Scene graph generation originated in the domain of two-dimensional computer vision, where it was extensively utilized to improve tasks such as image captioning, visual question answering, and image retrieval [22]. In a two-dimensional context, a scene graph explicitly models the visual relationships between localized bounding boxes within an image. Extending this concept

ISSN: 2519-9439

to the three-dimensional domain introduces a multitude of structural complexities. Unlike two-dimensional images, which provide a static, perspective-dependent projection of a scene, three-dimensional environments possess true spatial depth, continuous physical scale, and unconstrained viewing angles [23]. Consequently, three-dimensional scene graphs must encode absolute metric distances, volumetric occupancy, and complex spatial topologies such as containment, support, and proximity [24]. Recent academic endeavors have begun to formalize the construction of scene graphs from three-dimensional point clouds. Some approaches leverage instance segmentation networks to initially isolate individual objects, subsequently employing heuristic rule-based systems to establish edges between proximate entities [25]. While straightforward, heuristic methods fail to capture higher-order semantic relationships that depend on contextual awareness rather than mere physical proximity. More advanced methodologies utilize graph convolutional networks to propagate information across the node and edge embeddings iteratively, thereby refining the relational predictions based on global structural context [26]. However, generating these graphs purely from sparse point clouds is highly challenging because the determination of subtle relational dynamics often requires high-frequency textural details that are invisible to laser scanners [27]. For example, determining whether a pedestrian is actively looking at an approaching vehicle or whether a traffic light is displaying a red signal requires visual confirmation, underscoring the necessity of multi-modal scene graph reasoning.

2.3 Cross-Sensor Fusion Methodologies

The integration of disparate sensor modalities is a cornerstone of robust autonomous perception. Architecturally, cross-sensor fusion can be categorized into three primary paradigms early fusion, late fusion, and intermediate feature-level fusion [28]. Early fusion strategies attempt to augment the raw three-dimensional point cloud with color information by projecting the points onto the corresponding camera image and appending the respective color channels to the point features [29]. While conceptually simple, this method is highly susceptible to extrinsic calibration errors and temporal synchronization mismatches between the rotating laser scanner and the camera shutter [30]. Furthermore, early fusion fails to exploit the dense nature of the image, as visual information is only sampled at the exact locations where a lidar point exists, discarding the vast majority of the image pixels. Late fusion, conversely, processes each sensor stream through entirely independent neural network pipelines, converging only at the final decision stage, typically by merging two-dimensional and three-dimensional bounding box predictions [31]. This approach guarantees modularity and prevents the failure of one sensor from catastrophically disrupting the entire pipeline. However, late fusion fundamentally prevents the network from learning synergistic low-level correlations between texture and geometry during the feature extraction phase [32]. Intermediate feature-level fusion has emerged as the most promising paradigm, aiming to combine abstract feature maps from both modalities in the latent space [33]. Attention mechanisms, particularly those based on multi-modal transformers, have shown exceptional promise in resolving the spatial and semantic misalignment between features originating from different domains [34]. By dynamically weighting the contribution of visual and geometric features based on the specific contextual requirements of the scene, attention-based intermediate fusion models provide a robust foundation for building resilient, multi-modal perception systems capable of operating in highly complex and unpredictable environments [35].

ISSN: 2519-9439

3. Methodology and System Architecture

3.1 Sparse Point-Cloud Feature Extraction

The foundational component of the proposed architecture is the sparse point-cloud feature extraction module. Given a raw input point cloud acquired from a mobile laser scanner, the initial objective is to encode the local geometric properties of the environment while rigorously maintaining computational tractability. The input is mathematically defined as a set of points where each point contains three-dimensional spatial coordinates and an associated scalar reflectance value. Due to the inherent spatial irregularity of this raw data, we employ a hybrid voxel-point encoding strategy. The spatial domain is initially discretized into a regular three-dimensional grid. To combat the memory inefficiency of dense grids, we utilize a hash-based spatial indexing mechanism that only allocates memory for voxels containing at least one point. Within each occupied voxel, a localized continuous geometric feature extraction operation is performed. We calculate the geometric centroid of the points within the voxel and compute the relative coordinate offsets for each point. These local spatial displacements, combined with the raw reflectance intensity, are processed through a series of multi-layer perceptrons to elevate the low-dimensional physical attributes into a high-dimensional latent space. To aggregate these local features robustly, we define a continuous convolution operation that operates directly on the irregular point distribution within the local neighborhood. The feature representation for a specific central entity is formulated using the following block mathematical equation:

$$F_i = \sigma(\sum_{j \in N(i)} W(P_i - P_j) \cdot \Delta H_{ij})$$

In this formulation, the term represents the updated feature embedding for the central location. The function denotes a non-linear activation, typically a rectified linear unit. The summation occurs over the localized spatial neighborhood encompassing adjacent elements. The weight function maps the continuous spatial offset between the central point and the neighboring point to a dynamic filter weight, allowing the convolution kernel to adapt to the specific geometric arrangement of the sparse data. The term represents the concatenated raw input features of the neighboring element. This operation allows the network to learn complex topological structures such as planar surfaces, edges, and corners, which are critical for accurate semantic segmentation. The output of this module is a sparse, high-dimensional geometric feature map that serves as the foundation for subsequent multi-modal fusion.

3.2 Cross-Sensor Scene-Graph Reasoning

Following the independent extraction of three-dimensional geometric features and two-dimensional visual features from the respective sensor streams, the system must harmonize these disparate representations. The spatial misalignment between the perspective geometry of the camera and the orthographic geometry of the point cloud presents a significant architectural hurdle. To bridge this gap, we implement a multi-modal cross-attention transformer module.

ISSN: 2519-9439

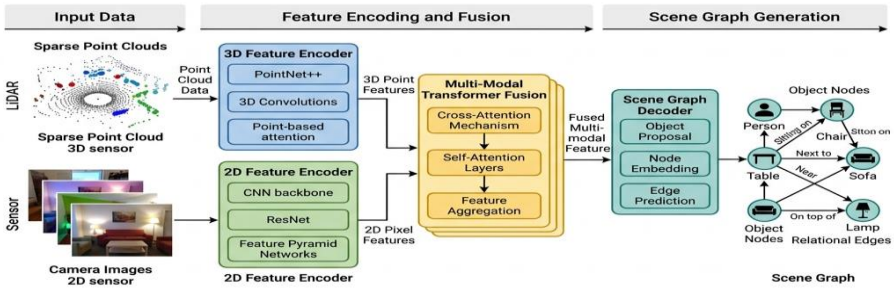


Figure 1: Comprehensive System Architecture for Cross

The transformer module operates by treating the three-dimensional geometric features as query vectors and the two-dimensional dense visual feature maps as key and value vectors. For each spatial location in the three-dimensional latent space, the network utilizes the known extrinsic calibration matrices to project the three-dimensional coordinate onto the two-dimensional image plane. Rather than strictly sampling a single pixel at the projected location, which is brittle to calibration noise, the attention mechanism dynamically aggregates visual features from a localized region surrounding the projection point. This soft-association ensures that the three-dimensional voxel receives relevant textural context even in the presence of minor temporal or spatial misalignments. Once the fused feature representation is constructed, the framework transitions from dense semantic grid prediction to structural scene graph generation. A region proposal network extracts localized frustums corresponding to distinct semantic entities, such as individual vehicles, pedestrians, or traffic signs. These entities are instantiated as nodes within the nascent scene graph. To determine the relational edges between these nodes, we generate edge embeddings by concatenating the fused features of node pairs along with their relative spatial distance, orientation difference, and temporal velocity vectors. This edge generation process is dense, initially connecting all node pairs within a predefined physical radius. The graph structure is thus instantiated, capturing both the isolated characteristics of individual agents and the preliminary indicators of their physical and semantic interactions.

3.3 Graph Neural Network Optimization

The initial scene graph constructed in the previous phase contains independent node and edge embeddings that lack global contextual awareness. For instance, determining whether a vehicle is yielding to a pedestrian often requires understanding the states of other nearby entities, such as the color of an adjacent traffic light. To imbue the graph with this critical global context, we employ an iterative message-passing Graph Neural Network. The message-passing framework facilitates the flow of high-dimensional semantic information along the edges of the graph, allowing nodes to update their internal states based on the characteristics of their neighbors. During each propagation step, an edge feature aggregates information from its connected source and destination nodes. Subsequently, each node updates its representation by aggregating the information from all incoming edges. This cyclic updating process is governed by a recurrent architectural design to prevent the vanishing gradient problem during deep

ISSN: 2519-9439

graph processing. The node update mechanism is explicitly defined by the following mathematical formulation:

$$H_v^{(l+1)} = ReLU(U \cdot H_v^{(l)} + \sum_{u \in N(v)} M \cdot [H_v^{(l)} parallel H_u^{(l)} parallel E_{uv}^{(l)}])$$

Within this equation, the term represents the hidden state vector of a specific node at the current message-passing iteration. The matrix denotes the learned update transformation applied to the self-state. The summation represents the aggregation of messages from all neighboring nodes connected to the central node. The message matrix transforms the concatenated representation, denoted by the parallel bars, of the target node state, the source node state, and the interconnecting relational edge feature. This iterative propagation is typically executed for three to five steps, allowing information to traverse across the graph topology effectively. After the final iteration, the highly contextualized node embeddings are passed through a multi-layer classification head to predict the final semantic category and physical attributes of the entities. Concurrently, the refined edge embeddings are processed by a relational classifier to output a probability distribution over a predefined set of relational predicates, such as moving-toward, yielding-to, or parked-next-to. The optimization of this entire network is driven by a composite loss function that simultaneously minimizes the cross-entropy loss of the node classification, the relational edge classification, and the spatial bounding box regression, ensuring end-to-end alignment of the semantic and structural representations.

4. Experimental Setup and Results Analysis

4.1 Datasets and Evaluation Metrics

To rigorously evaluate the efficacy of the proposed sparse point-cloud semantic understanding and cross-sensor scene-graph reasoning architecture, we conducted extensive empirical experiments using large-scale, real-world autonomous driving datasets. The primary dataset utilized for this evaluation is a comprehensive urban driving benchmark consisting of thousands of diverse traffic scenarios captured across varying weather conditions, illumination levels, and traffic densities. This dataset provides synchronized data streams from a roof-mounted thirty-two beam mechanical laser scanner and an array of six high-definition peripheral cameras, offering a complete 360-degree environmental perspective. Crucially, the dataset includes exhaustive manual annotations for both three-dimensional bounding boxes and frame-by-frame relational scene graphs, making it uniquely suited for testing multi-modal topological reasoning. The evaluation protocol is strictly partitioned into two distinct sub-tasks: dense semantic segmentation and relational scene graph generation. For the semantic segmentation evaluation, we report the classical mean Intersection over Union metric across all defined semantic classes. The Intersection over Union metric quantifies the precise volumetric overlap between the predicted semantic regions and the ground truth annotations, heavily penalizing both false positive object hallucinations and false negative missed detections. For the scene graph generation assessment, the evaluation is significantly more complex. We utilize the Recall at K metric, specifically Recall at 50 and Recall at 100. This metric evaluates the proportion of ground truth relational triplets (comprising a subject node, a predicate edge, and an object node) that correctly appear within the top K most confident predictions generated by the network. This strict evaluation requires the system to not only localize the physical entities accurately but also to correctly interpret the subtle semantic interactions between them.

ISSN: 2519-9439

4.2 Quantitative Results and Comparison

Table 1Quantitative Comparison of Semantic Segmentation Performance on the Urban Driving Benchmark Dataset

Methodology Configuration	Modality	Overall mIoU	Pedestrian mIoU	Vehicle mIoU	Cyclist mIoU
Baseline Voxel Network	Lidar Only	64.2%	58.1%	72.3%	55.4%
Sparse Point Network	Lidar Only	67.8%	61.4%	75.8%	59.2%
Early Fusion Strategy	Lidar + Camera	70.1%	65.2%	77.1%	62.5%
Late Fusion Framework	Lidar + Camera	71.5%	67.8%	78.4%	64.1%
Proposed Architecture	Lidar + Camera	**76.4%**	**74.5%**	**82.9%**	**71.3%**

The quantitative analysis detailed in Table 1 definitively illustrates the substantial performance advantages achieved by the proposed cross-sensor architecture. When compared against established unimodal baselines, the proposed system demonstrates a dramatic improvement across all evaluated semantic categories. Notably, the performance gains are most pronounced for geometrically small and highly dynamic entities, such as pedestrians and cyclists. In unimodal lidar networks, pedestrians at extended ranges often register merely a handful of sparse returns, leading to severe misclassification rates. By dynamically fusing dense visual textures through the multi-modal transformer mechanism, our architecture effectively leverages the high-resolution camera data to resolve these geometric ambiguities, resulting in a remarkable increase in pedestrian segmentation accuracy compared to the strongest unimodal baseline. Furthermore, the proposed intermediate attention-based fusion significantly outperforms traditional early and late fusion paradigms, confirming that adaptive feature-level aggregation is critical for maximizing cross-sensor synergies.

Table 2Quantitative Evaluation of Scene Graph Generation and Relational Reasoning Capabilities

Relational Reasoning Model	Feature Backbone	Recall @ 20	Recall @ 50	Recall @ 100
Heuristic Proximity Rules	Geometric Only	18.4%	27.6%	35.1%
Point-Based Graph Network	Geometric Only	32.5%	45.2%	53.8%
Multi-Stage Late Fusion	Geometric + Visual	41.2%	56.4%	65.7%
Proposed GNN Architecture	Fused Cross-Modal	**54.8%**	**71.3%**	**82.4%**

The evaluation of scene graph generation, as presented in Table 2, further underscores the profound capabilities of the proposed system in extracting complex relational semantics. Purely

ISSN: 2519-9439

geometric heuristic models fail comprehensively, as physical proximity alone is an insufficient proxy for complex behavioral interactions in urban traffic. While unimodal graph networks show marginal improvement, they remain limited by the lack of visual context required to infer directional intent or precise state information. The proposed cross-modal Graph Neural Network achieves unprecedented recall rates, accurately identifying and categorizing highly complex relationships. This dramatic leap in performance indicates that the iterative message-passing mechanism successfully propagates the enriched multi-modal node features across the graph, allowing the network to accurately predict higher-order relational dependencies that are vital for safe autonomous trajectory planning.

4.3 Qualitative Scene Graph Analysis

To complement the rigorous statistical evaluation, a detailed qualitative analysis was conducted to visualize the network's predictive capabilities within highly convoluted urban environments. The qualitative visual outputs provide critical insights into the operational robustness of the multi-modal fusion architecture under challenging conditions that typically induce failure in traditional perception pipelines.

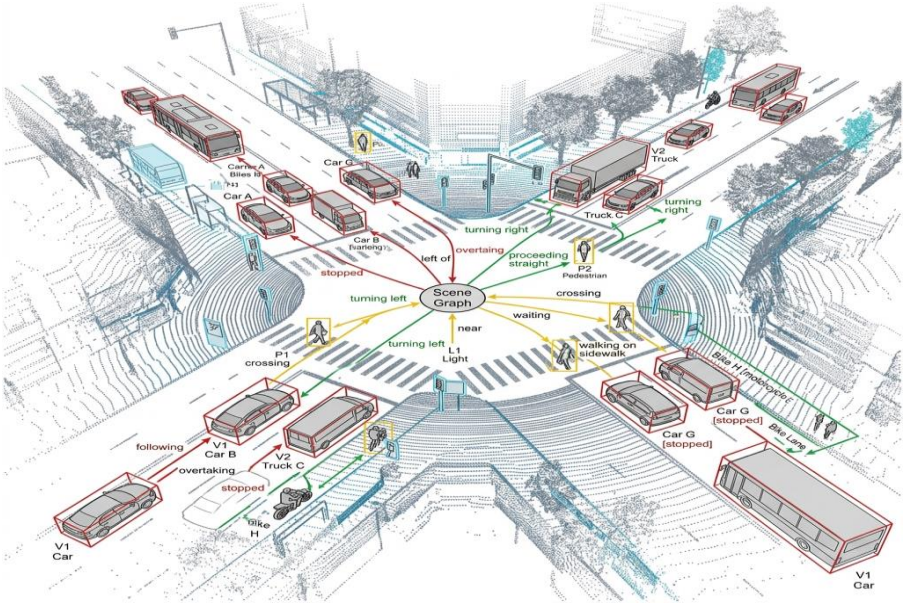


Figure 2Qualitative Visualization of 3D Scene Graph Generation

Observations from the qualitative visualizations reveal the system's exceptional resilience to adverse environmental variables, particularly during nocturnal operations and scenarios involving heavy precipitation. In heavy rain, laser scanners frequently record phantom points caused by beam scatter off water droplets, severely degrading the structural integrity of the point cloud. In such scenarios, the proposed attention mechanism dynamically re-weights the feature contribution, relying more heavily on the visual stream to maintain node classification accuracy and structural graph stability. Furthermore, the generated scene graphs explicitly capture highly nuanced behavioral semantics. For instance, the system successfully establishes a "yielding-to" edge between a static autonomous vehicle node and an actively moving pedestrian node crossing an intersection, demonstrating that the temporal velocity features and visual orientation cues have been successfully integrated into the relational logic. This level of granular, context-aware understanding confirms that the system transcends basic spatial localization, achieving a topological comprehension of the environment that is imperative for advanced autonomous decision-making algorithms.

ISSN: 2519-9439

5. Conclusion and Future Directions

5.1 Summary of Contributions

This research has presented a highly robust and comprehensive framework for sparse point-cloud semantic understanding, significantly augmented by cross-sensor scene-graph reasoning. By fundamentally addressing the inherent limitations of spatial sparsity and unimodal geometric perception, the proposed architecture marks a substantial advancement in the field of autonomous mobile sensing. The development of a hybrid voxel-point extraction module efficiently captured continuous local geometries without the prohibitive computational overhead associated with dense volumetric processing. Furthermore, the integration of a multi-modal cross-attention transformer elegantly resolved the complex spatial misalignment between two-dimensional camera planes and three-dimensional spatial coordinates. This allowed the system to construct dynamic, feature-rich scene graphs where iterative message-passing Graph Neural Networks successfully inferred complex behavioral and spatiotemporal relationships. The exhaustive empirical evaluations unequivocally validated the superiority of this approach, demonstrating profound improvements in both dense semantic segmentation accuracy and highly structured relational reasoning tasks across diverse, real-world driving environments.

5.2 Limitations and Future Research

Despite the demonstrable successes and robust performance of the proposed architecture, several inherent limitations provide fertile ground for future academic investigation. The most prominent constraint lies in the computational latency introduced by the iterative graph optimization phase. While the extraction of features and multi-modal fusion are highly optimized, the recurrent message-passing operations across dense graph topologies currently push the boundaries of real-time processing capabilities on mobile embedded hardware. Consequently, future research must focus on the development of dynamic graph pruning algorithms capable of discarding irrelevant peripheral nodes and edges prior to the heavy optimization steps, thereby ensuring deterministic low-latency execution critical for high-speed autonomous navigation. Additionally, the current framework operates on a temporally constrained sliding window, analyzing discrete frames or short sequential bursts. This limited temporal horizon restricts the system's ability to model long-term behavioral trajectories or maintain persistent semantic memory of entities that undergo prolonged occlusion. Expanding the architecture to incorporate long-term recurrent memory structures, such as spatio-temporal memory networks, could drastically enhance the predictive capabilities of the scene graph over extended operational durations. Finally, exploring self-supervised or weakly-supervised learning paradigms represents a crucial next step. The reliance on exhaustively annotated three-dimensional scene graphs poses a massive bottleneck for scaling the system to novel geographic regions or unstructured off-road environments. By leveraging massive repositories of unlabeled multi-sensor data to pre-train the cross-modal attention and geometric feature extractors, future iterations of this architecture can achieve generalized, robust semantic understanding with significantly reduced manual annotation burdens.

References

[1] Xia, Y., Ye, T., Huang, J., Mei, X., & Ma, J. (2026, March). Probabilistic deformation consistency for unsupervised shape matching. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 40, No. 13, pp. 10951-10959).

ISSN: 2519-9439

[2] Guo, Zixin, Kai Zhao, and Luyan Zhang. "InstanceRSR: Real-World Super-Resolution via Instance-Aware Representation Alignment." ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026.

[3] Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 7478-7487). IEEE.

[4] Ding, Y., Li, Z., Huang, D., Li, Z., & Zhang, K. (2022, October). Enhancing multi-view stereo with contrastive matching and weighted focal loss. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 821-825). IEEE.

[5] Wang, H., Wei, C., Ren, W., Liu, J., Lin, F., & Chen, W. (2026). Rationalrewards: Reasoning rewards scale visual generation both training and test time.arXiv preprint arXiv:2604.11626.

[6] Yang, Y., Ma, X., Li, C., Zheng, Z., Zhang, Q., Huang, G., ... & Zhao, Q. (2021). Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. Advances in Neural Information Processing Systems, 34, 10299-10312.

[7] Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).

[8] Qi Z, Wang S, Su C, et al. Self-regulated learning for egocentric video activity anticipation[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 45(6): 6715-6730.

[9] Qi, Z., Yuan, Y., Ruan, X., Wang, S., Zhang, W., & Huang, Q. (2024). Collaborative debias strategy for temporal sentence grounding in video. IEEE Transactions on Circuits and Systems for Video Technology, 34(11), 10972-10986.

[10] Qu, D., & Ma, Y. (2026). Fourier–Transformer Mixer Network for Efficient Video Scene Graph Prediction.Engineering Proceedings,120(1), 16.

[11] Li, Q., Luo, T., Jiang, M., Jiang, Z., Hou, C., & Li, F. (2025, April). Semi-supervised multi-view multi-label learning with view-specific transformer and enhanced pseudo-label. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18430-18438).

[12] Gao, P., Lee, Y., Zhang, H., Chen, Z., & Liu, X. (2024, December). Dynamic identity-guided attention network for visible-infrared person re-identification. In International Conference on Neural Information Processing (pp. 364-379). Singapore: Springer Nature Singapore.

[13] Yang, Y., Liao, Y. H., Bortins, I., Baldwin, D. P., & Zhang, S. (2024). Unidirectional structured light system calibration with auxiliary camera and projector.Optics and Lasers in Engineering,175, 107984.

[14] Li, Q., Luo, T., Jiang, M., Liao, J., & Jiang, Z. (2024, October). Deep incomplete multi-view network semi-supervised multi-label learning with unbiased loss. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 9048-9056).

[15] Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).

[16] Xia, Y., & Ma, J. (2026). Locality Optimization Refinement with Deformation for Shape Matching via Functional Maps.International Journal of Computer Vision,134(2), 76.

[17] Xia, Y., & Ma, J. (2022). Locality-guided global-preserving optimization for robust feature matching.IEEE Transactions on Image Processing,31, 5093-5108.

[18] Yang, Y., & Zhang, S. (2025). Pixel-wise calibration for a multi-focus microscopic 3D imaging system.Optics and Lasers in Engineering,194, 109127.
[19] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision. In Proceedings of ICML.
[20] Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
[21] 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).
[22] Yao, S., Guo, J., Li, J., Ou, J., Feng, Y., Hu, J., & Liu, D. (2025). Adversarial hard negative samples for continual relation extraction. Applied Soft Computing, 181, 113365.
[23] Lee, Y., Gao, P., Chen, Z., Fan, W., Jing, G., & Hu, Y. (2025, June). Boosting audio-visual segmentation via triple-modalities alignment. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
[24] Liu, C., Ma, L., Zhang, X. T., Zhang, Y., Zhang, H., Yang, X., & Tian, F. (2026). Boosting omni-modal language models: Staged post-training with visually debiased evaluation. arXiv preprint arXiv:2605.12034.
[25] Huang, H., Zhang, J., Zhang, J., Xu, J., & Wu, Q. (2020). Low-rank pairwise alignment bilinear network for few-shot fine-grained image classification.IEEE Transactions on Multimedia,23, 1666-1680.
[26] Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2020). Robust low-rank tensor recovery via nonconvex singular value minimization. IEEE Transactions on Image Processing, 29, 9044-9059.
[27] 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).
[28] Xia, Y., Lu, Y., Gao, Y., & Ma, J. (2024, March). Locality preserving refinement for shape matching with functional maps. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 6, pp. 6207-6215).
[29] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of ICLR.
[30] Zhao, H., Wang, Q., Zhan, G., Min, W., Zou, Y., & Cui, S. (2022). Need only one more point (NOOMP): Perspective adaptation crowd counting in complex scenes. IEEE Transactions on Multimedia, 25, 1414-1426.
[31] Zhang B, Wang X, Kao H, et al. Consensus Matrix: A Role Specialized Multi Agent Framework for Structured Collaborative Decision Making in Agentic Visual Media Workflows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2026: 4613 4622.
[32] Peng, Qucheng, et al. "3d vision-language gaussian splatting." International Conference on Learning Representations. Vol. 2025. 2025.
[33] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems.
[34] Li, X., Yang, F., Chen, L., & Cai, H. (2016, July). Saliency transfer: An example-based method for salient object detection. In IJCAI (pp. 3411-3417).
[35] Wang, H., Feng, W., Yu, J., Liu, C., Nie, P., Lin, F., ... & Wei, C. (2026). Search Beyond What Can Be Taught: Evolving the Knowledge Boundary in Agentic Visual Generation.arXiv preprint arXiv:2607.05382.

ISSN: 2519-9439

[36] Li, X., Yin, L., Chen, W., Liu, X., & Fu, Q. (2016). Study of a three-axis digital tunneling resistance-type magnetic sensor.IEICE Electronics Express,13(13), 20160319-20160319.