Sparse-Aware Vector Quantization for Bandwidth-Efficient Collaborative 3D Semantic Occupancy Prediction

Feng Li[1], Chaokun Zhang[2⋆], and Gong Chen[1]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, China
[2] School of Cybersecurity, Tianjin University, Tianjin, China
{fengli0116, zhangchaokun, gongchen01}@tju.edu.cn

Abstract. Collaborative perception extends single-agent perception by enabling multiple vehicles to exchange complementary perceptual information. However, it introduces an inherent trade-off between perception gain and communication overhead, which is particularly severe for 3D semantic occupancy prediction that relies on fine-grained spatial structures. Existing methods typically compress 3D features into 2D, causing severe spatial information loss, or transmit dense 3D representations, hindering real-world deployment. To overcome these limitations, we propose a bandwidth-efficient collaborative **V**ector **Q**uantization **S**emantic **O**ccupancy **P**rediction (VQSOP) framework. VQSOP employs a Sparse-Aware Vector Quantization (SAVQ) mechanism that exploits 3D scene sparsity to compactly encode informative regions, drastically reducing communication overhead while preserving complete geometric context. Furthermore, to enhance structural consistency and feature continuity, we design a Dual-Branch Adaptive Spatial Refinement (ASR) module that dynamically fuses local high-frequency details with broad contextual semantics. Extensive experiments demonstrate that our approach achieves state-of-the-art performance while reducing communication volume by up to 82×.

Keywords: Collaborative perception · 3D occupancy prediction · Autonomous driving

1 Introduction

Reliable environmental perception forms the foundation of autonomous driving systems, as it directly determines the safety and effectiveness of downstream planning and control modules [22, 23, 43]. By leveraging Vehicle-to-Everything (V2X) communication, collaborative perception enables multiple agents to exchange complementary information and extend the effective perception range [11, 12, 40, 44]. This paradigm overcomes the inherent limitations of traditional

⋆ Corresponding author.

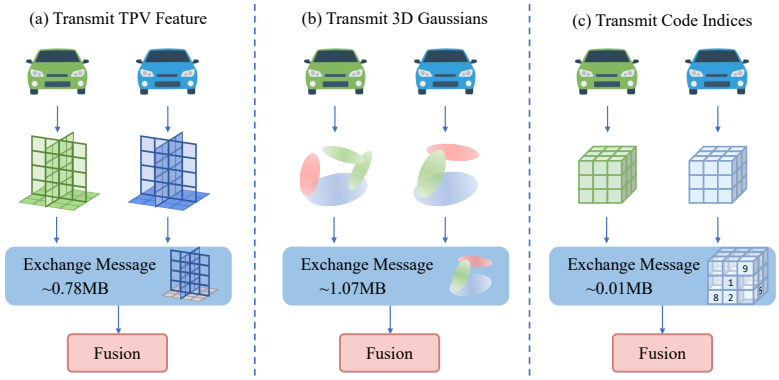

Fig. 1: Comparison of shared feature representations. (a) Transmitting TPV features loses geometric details and complicates spatial alignment. (b) Transmitting 3D Gaussians tightly couples prediction performance with communication bandwidth, as the number of Gaussians directly impacts accuracy. (c) Our method transmits more critical perception information via compact code index messages, drastically reducing the communication volume while preserving essential semantic and geometric information for effective collaborative fusion.

single-agent perception systems caused by restricted sensor fields of view and severe occlusions [10,37,43], thereby significantly enhancing robustness in complex and dynamic traffic environments.

Previous collaborative perception research predominantly focused on 3D object detection and Bird's-Eye-View (BEV) segmentation tasks [7,17,28,33,34,38]. These conventional tasks simplify scene perception by omitting crucial 3D semantic details. Furthermore, BEV-based features inherently suffer from information loss due to height compression. Such limitations restrict the ability to identify irregularly shaped or heavily occluded objects, ultimately hindering holistic scene understanding and downstream decision-making. Recently, 3D semantic occupancy prediction has garnered significant attention for its ability to provide detailed 3D semantic and geometric information by utilizing fine-grained voxel-based representations [2, 9, 14–16, 35].

Despite its tremendous potential to offer a richer understanding of the environment, 3D semantic occupancy prediction remains largely under-explored in collaborative settings [36]. Fundamentally, collaborative perception involves an inherent trade-off between perception gain and communication overhead. This conflict is particularly severe for 3D occupancy prediction, where high-dimensional spatial structures make the direct transmission of dense features impractical under limited communication bandwidth. CoHFF [26] proposed an early collaborative semantic occupancy prediction framework, demonstrating the immense potential of V2X feature fusion for semantic occupancy prediction. However, CoHFF adopts a tri-perspective view (TPV) representation that re-

lies on explicit depth supervision to preserve spatial consistency. Furthermore, to reduce communication volume, it only transmits orthogonal plane features, resulting in inevitable information loss (Fig. 1(a)). In an effort to retain complete 3D structural details without relying on 2D projections, recent methods have attempted to introduce 3D Gaussians [3]. Although these approaches preserve richer geometric expressiveness, prediction accuracy directly depends on the number of transmitted Gaussians. This tightly couples performance with communication bandwidth, fundamentally restricting their practical deployment in bandwidth-constrained collaborative perception systems (Fig. 1(b)).

In this paper, we propose a novel collaborative Vector Quantization Semantic Occupancy Prediction (VQSOP) framework that significantly improves the perception-communication trade-off in multi-agent systems. VQSOP introduces a Sparse-Aware Vector Quantization (SAVQ) mechanism that exploits inherent spatial sparsity and local feature redundancy in autonomous driving scenarios to selectively focus on informative regions. Instead of transmitting dense continuous features, this mechanism directly quantizes the selected high-dimensional feature representations into compact discrete codes defined in a shared codebook, transmitting only the corresponding indices. This quantization-driven design effectively reduces communication overhead while preserving essential 3D geometric context. Furthermore, to enhance semantic perception capabilities, we design a dual-branch Adaptive Spatial Refinement (ASR) module. The module consists of a local branch for preserving fine-grained geometric details and a context branch for modeling long-range semantic dependencies. These complementary features are adaptively fused via spatially adaptive weighting, enabling the model to reconstruct coherent and semantically consistent volumetric representations with high fidelity (Fig. 1(c)).

Extensive experiments validate the effectiveness of our proposed approach. Specifically, under the single-agent setting, VQSOP outperforms the strongest baseline by 4.41% in mIoU and 2.64% in IoU. Furthermore, in the collaborative setting, it surpasses state-of-the-art methods by 4.10% and 0.92% in mIoU and IoU, respectively. Notably, VQSOP reduces communication volume by up to 82× while delivering superior collaborative perception performance.

Our main contributions can be summarized as follows:

- We propose VQSOP, a novel collaborative semantic occupancy prediction framework. It compresses features into discrete code indices for transmission, effectively preserving perception-relevant information.
- We design a SAVQ mechanism to compactly encode key informative regions. Additionally, we introduce an ASR module to dynamically aggregate spatial features and enhance the overall semantic perception capabilities.
- We conduct comprehensive experiments to validate our approach, demonstrating that VQSOP achieves state-of-the-art performance and establishes a highly superior perception-communication trade-off.

The code is available at `https://github.com/cheerfulli/VQSOP`.

2 Related Work

2.1 3D Semantic Occupancy Prediction

3D semantic occupancy prediction has attracted growing interest in recent years, as it provides a holistic representation of driving environments by predicting both occupancy and semantic labels for all voxels within a predefined spatial range. Existing methods can be broadly categorized based on their input sensory modalities. LiDAR-based approaches leverage the precise geometric measurements of 3D point clouds to generate high-quality voxel predictions [6, 42, 46]. Meanwhile, vision-centric approaches have dominated the mainstream due to their rich semantic context and remarkable cost-effectiveness. SurroundOcc [35] utilized spatial attention to lift 2D features into 3D space in a multi-scale manner, employing multi-frame point clouds to construct dense occupancy ground truths. TPVFormer [14] proposed a tri-perspective view representation for describing 3D scenes in semantic occupancy prediction. OccFormer [45] designed a dual-path Transformer network to process 3D volume features efficiently. SGN [21] presented a dense–sparse–dense architecture that adaptively identifies sparse seed voxels and incorporates hybrid guidance to promote the convergence of semantic propagation. GaussianFormer [15] introduced an object-centric approach that describes 3D scenes using sparse 3D semantic Gaussians. Liu et al. [19] proposed a fully sparse occupancy network that reconstructs sparse geometry with a sparse voxel decoder and predicts semantic occupancy using sparse queries. Tang et al. [29] leveraged a lossless sparse latent representation to achieve efficient semantic occupancy prediction. OPUS [32] formulates occupancy prediction as a direct set prediction problem, leveraging learnable queries to predict occupied locations and classes. Despite the remarkable progress achieved by these vision-centric semantic occupancy prediction methods, they are primarily designed for single-agent systems. Directly extending these high-dimensional 3D representations to multi-agent collaborative scenarios introduces severe communication bottlenecks.

2.2 Collaborative Perception

Collaborative perception paradigms can be broadly classified into early, late, and intermediate fusion. Early fusion methods [5, 27] share raw sensor data between connected agents to expand the perceptive field, but consume excessive communication bandwidth, severely limiting their applicability in real-time. Late fusion approaches [1, 25] share and fuse the final detection outputs from individual vehicles. Although this paradigm is highly bandwidth-efficient, it is extremely sensitive to the individual detection performance of each agent, where a single localized error can easily degrade the final consensus. To strike an optimal balance between communication overhead and perception accuracy, recent efforts have predominantly focused on intermediate fusion, where agents share processed intermediate feature representations. V2X-ViT [40] introduces a heterogeneous

multi-agent attention module to fuse information from vehicles and infrastructure. Where2comm [11] proposes a novel spatial confidence-aware communication strategy that focuses on key perceptual regions, enabling performance improvement with reduced communication overhead. CORE [31] utilizes LiDAR measurements and adopts spatial-channel compression along with attention-based collaboration for efficient scene reconstruction. CoBEVT [38] designs a local-global sparse attention mechanism to capture complex spatial interactions across different views and agents. CoBEVFusion [24] investigates multi-modal fusion by introducing a Dual Window-based Cross-Attention module, which integrates camera and LiDAR features through CNN-based inter-agent interaction. WhisperNet [4] proposes a novel receiver-centric global coordination paradigm that jointly optimizes feature transmission in the spatial and channel dimensions across agents. However, these intermediate fusion strategies are predominantly tailored for 3D object detection or 2D BEV segmentation. Directly applying them to transmit dense, high-dimensional features for 3D semantic occupancy prediction incurs an unacceptable bandwidth burden.

3 Method

3.1 Problem Formulation

We model the collaborative perception system as a communication graph $\mathbb{G} = (\mathcal{A}, \mathcal{E})$, where $\mathcal{A}$ is the set of agents, and $\mathcal{E}$ represents the existing communication links between two agents. For an agent $i \in \mathcal{A}$, its connected neighbors are defined as $\Omega_i = \{j \mid (i,j) \in \mathcal{E}, j \neq i\}$. Let $\mathbf{X}_i$ denote the multi-view RGB images captured by the surround cameras mounted on agent i. The holistic surrounding environment is represented as 3D voxels with one-hot embeddings $\mathbf{O} \in \mathbb{R}^{X \times Y \times Z \times C}$, where X, Y, and Z represent the spatial dimensions, and C is the number of semantic categories. For each agent i, $\mathbf{O}_i$ denotes the predicted occupancy of the voxels, while $\mathbf{O}_i^*$ represents the corresponding ground truth. The collaborative semantic occupancy prediction task is formulated as a constrained optimization problem:

$$\begin{aligned} \max_{\theta, \mathcal{P}} & \sum_{i \in \mathcal{A}} g\Big(\Phi_\theta\big(\mathbf{X}_i, \{\mathcal{P}_{j \to i}\}_{j \in \Omega_i}\big), \mathbf{O}_i^*\Big), \\ \text{s.t.} \quad & \sum_{i \in \mathcal{A}} \sum_{j \in \Omega_i} b(\mathcal{P}_{j \to i}) \leq B, \end{aligned} \tag{1}$$

where Φ_θ denotes the collaborative perception network parameterized by θ, and $g(\cdot)$ is the prediction evaluation metric. $\mathcal{P}_{j \to i}$ is the message transmitted from agent j to agent i, and $b(\cdot)$ measures the communication cost of the collaborative messages. Ultimately, the optimization objective is to maximize the overall perception effectiveness under the communication upper bound $B \in \mathbb{R}^+$.

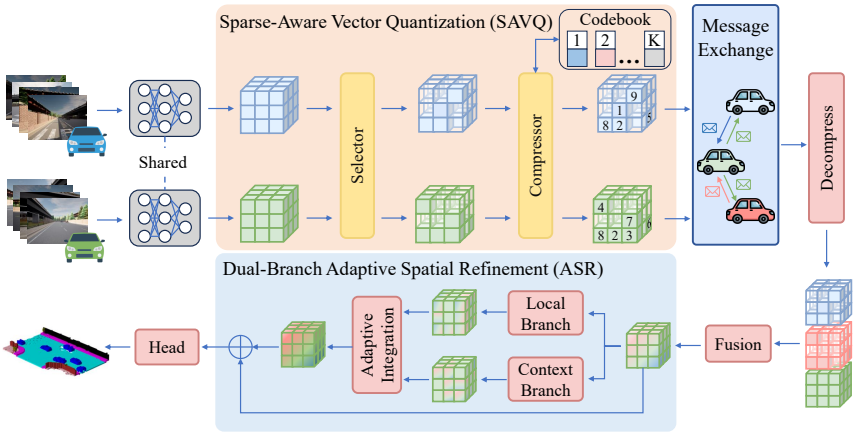

Fig. 2: Overall architecture of the proposed VQSOP framework. The pipeline consists of three main stages: (1) SAVQ mechanism, which compresses dense 3D spatial features into discrete code indices for bandwidth-efficient transmission; (2) message decompression and fusion, where received neighbor messages are reconstructed and spatially aggregated with the ego agent's local representation; and (3) ASR module, which dynamically integrates local geometric details and broad contextual semantics via spatially adaptive weighting to yield the final 3D semantic occupancy prediction.

3.2 Overall Architecture

Fig. 2 illustrates the overall pipeline of our proposed VQSOP framework. Each agent utilizes a shared backbone to process multi-view RGB images and metadata, lifting them into continuous 3D occupancy features. To overcome communication bandwidth bottlenecks, these local features are processed by the SAVQ mechanism before transmission. Specifically, a selector isolates informative spatial regions, and a compressor maps these continuous features into highly compact discrete indices by querying a learnable codebook. These discrete indices are then transmitted through the V2X message exchange network. Upon reception, the message decompression module queries the shared codebook to reconstruct the continuous 3D neighbor features from the received indices. Subsequently, the collaborative fusion module spatially aggregates these reconstructed features with the ego agent's local representation. To further enhance the semantic perception capabilities, the fused features are fed into the ASR module. The ASR utilizes a local branch to extract fine-grained geometric details and a context branch to capture broad contextual semantics. These diverse spatial features are dynamically integrated via spatially adaptive weighting and a residual connection to reconstruct high-fidelity structural details. Finally, the refined features are passed to a task-specific prediction head to generate the ultimate 3D semantic occupancy prediction.

3.3 Sparse-Aware Vector Quantization

In collaborative perception, 3D voxel features require higher communication bandwidth than 2D BEV representations, making the direct transmission of dense 3D volumes a major bottleneck for real-world deployment. Nevertheless, we observe that 3D driving scenes naturally exhibit high spatial sparsity, where the vast majority of voxels are empty, implying that only a small fraction contains informative semantic and geometric cues. Therefore, to effectively filter key regions and compress high-dimensional features, we design the SAVQ mechanism. As shown in Fig. 3, this mechanism selectively retains informative regions and maps the continuous features into a compact discrete codebook representation, thereby significantly reducing communication costs.

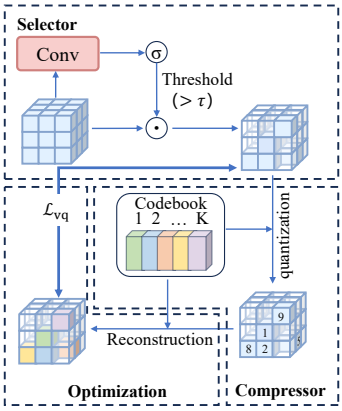

Fig. 3: Architecture of the SAVQ mechanism.

Sparse-Aware Selector. For the j-th agent in the collaborative network, given a dense 3D feature volume $\mathbf{V}_j \in \mathbb{R}^{X\times Y\times Z\times C}$ extracted by the shared backbone, we employ a lightweight convolutional module to predict a spatial confidence map $\mathbf{S}_j \in [0,1]^{X\times Y\times Z}$, which indicates the importance of each voxel. A binary sparse mask $\mathbf{M}_j$ is then generated by applying a predefined confidence threshold τ:

$$\mathbf{M}_j(x,y,z) = \begin{cases} 1, & \text{if } \mathbf{S}_j(x,y,z) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

By applying this mask to the original feature volume, we obtain the filtered feature map $\mathbf{F}_j$:

$$\mathbf{F}_j = \mathbf{V}_j \odot \mathbf{M}_j, \tag{3}$$

where $\odot$ denotes element-wise multiplication broadcasted along the channel dimension. This step ensures that the subsequent quantization and transmission are strictly focused on geometry-aware and semantic-rich areas.

Codebook-Based Compressor. To efficiently compress high-dimensional voxel semantic data, we introduce a voxel semantic compression scheme using learnable codebooks. Inspired by [13, 30], we approximate high-dimensional feature vectors using the most relevant codes from a task-driven codebook. Consequently, only integer code indices need to be transmitted, rather than complete feature vectors composed of floating-point numbers. This codebook consists of a set of prototype codes that are optimized to effectively approximate the diverse perceptual features present in driving scenes. Let $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$ denote the learnable codebook, where K is the dictionary size and $\mathbf{c}_k$ is the k-th prototype code. For each valid spatial position (x, y, z) selected by the mask (i.e.,

$\mathbf{M}_j(x, y, z) = 1$), we map the continuous feature $\mathbf{F}_j(x, y, z)$ to a discrete code index via a nearest-neighbor lookup:

$$I_j(x, y, z) = \arg \min_{k \in \{1,2,\ldots,K\}} \|\mathbf{F}_j(x, y, z) - \mathbf{c}_k\|_2. \tag{4}$$

During transmission, the j-th agent only needs to broadcast the discrete indices I_j corresponding to the critical regions. The bandwidth required to transmit an index is merely $\log_2(K)$ bits (e.g., 8 bits for $K = 256$), which achieves a massive compression ratio compared to the original dense volume.

Codebook Optimization. To ensure the learnable codebook accurately represents the continuous feature space, we optimize it by minimizing the reconstruction error. Let $\hat{\mathbf{F}}_j$ denote the quantized sparse feature map, where each valid position is mapped back to its corresponding prototype: $\hat{\mathbf{F}}_j(x, y, z) = \mathbf{c}_{I_j(x,y,z)}$, and the filtered empty regions remain zero. To optimize the codebook, we define the quantization loss across all agents as:

$$\mathcal{L}_{vq} = \sum_{j \in \mathcal{A}} \sum_{x,y,z} \left\| \mathbf{F}_j(x, y, z) - \hat{\mathbf{F}}_j(x, y, z) \right\|_2^2. \tag{5}$$

3.4 Message Decompression and Collaborative Fusion

Upon receiving the compact messages from neighboring agents, the ego agent must decode these discrete representations and spatially aggregate them to construct a holistic 3D scene representation.

Message Decompression. Given the transmitted integer indices from a neighboring agent j, the ego agent reconstructs the quantized feature volume by querying the shared learnable codebook $\mathcal{C}$. Each discrete index corresponds to a prototype vector in the codebook, enabling the recovery of continuous feature representations from compact messages. The reconstructed volume is denoted as $\hat{\mathbf{F}}_j \in \mathbb{R}^{X \times Y \times Z \times C}$. For each valid spatial location indicated by $I_j(x, y, z)$, the feature is obtained through a direct lookup operation, $\hat{\mathbf{F}}_j(x, y, z) = \mathbf{c}_{I_j(x,y,z)}$. The remaining background regions that were pruned during transmission are padded with zeros to maintain a consistent spatial tensor structure. The resulting feature volume is then spatially aligned and fused with the ego vehicle's local representation in subsequent stages.

Collaborative Fusion. To construct a holistic 3D scene representation, the ego agent aggregates its own local dense feature volume $\mathbf{V}_i$ with all reconstructed neighbor features to generate the comprehensive collaborative representation:

$$\hat{\mathbf{F}}_{fused} = \Psi_{fuse}\Big(\mathbf{V}_i, \{\hat{\mathbf{F}}_j\}_{j \in \Omega_i}\Big), \tag{6}$$

where $\Psi_{fuse}(\cdot)$ denotes the feature fusion operator (e.g., an attention-based spatial aggregator) that effectively integrates multi-agent perspectives. This collaborative fused volume $\hat{\mathbf{F}}_{fused}$ subsequently serves as the direct input to the downstream refinement module.

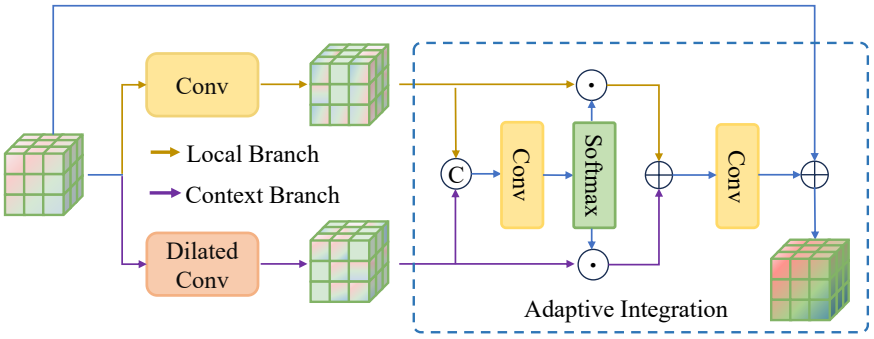

Fig. 4: Architecture of the ASR module. It dynamically aggregates local geometric details and broad contextual semantics through a parallel dual-branch design with spatially adaptive weighting.

3.5 Dual-Branch Adaptive Spatial Refinement

While collaborative feature fusion enhances spatial awareness, it may blur fine-grained geometric boundaries during aggregation, and long-range contextual dependencies are not fully captured by standard convolutions, hindering precise semantic occupancy prediction. Therefore, we propose an ASR module to enhance both local geometric details and broad contextual semantics. As shown in Fig. 4, the ASR module adopts a parallel dual-branch design: a local branch with standard 3D convolutions for fine-scale details, and a context branch with dilated convolutions for broader dependencies. Their outputs are fused via spatially adaptive weighting for dynamic refinement.

Local Branch. The local branch focuses on enhancing fine-grained geometric structures within the 3D feature volume. Specifically, it employs stacked standard 3D convolutional layers with small receptive fields to enhance local spatial continuity and boundary sharpness. This helps recover detailed geometric patterns that may be weakened during multi-agent aggregation, ultimately generating the local representation $\mathbf{F}_{\text{local}} \in \mathbb{R}^{X \times Y \times Z \times C}$.

Context Branch. The context branch aims to capture long-range semantic dependencies across the scene. Specifically, it employs 3D dilated convolutions to efficiently enlarge the spatial receptive field, which enables the aggregation of broader contextual information without sacrificing resolution, facilitates consistent semantic reasoning over spatially distant regions, and produces the context representation $\mathbf{F}_{\text{context}} \in \mathbb{R}^{X \times Y \times Z \times C}$.

Adaptive Integration. Rather than using a naive addition to combine the multi-scale features, we design a spatially adaptive integration mechanism to dynamically fuse the two branches. We concatenate $\mathbf{F}_{\text{local}}$ and $\mathbf{F}_{\text{context}}$ and pass them through a lightweight transformation followed by a Softmax activation. This generates two complementary spatial weight maps, $\mathbf{W}_{\text{local}}, \mathbf{W}_{\text{context}} \in$

$\mathbb{R}^{X\times Y\times Z\times 1}$, which represent the specific demand of each voxel for either boundary details or contextual inpainting:

$$[\mathbf{W}_{\text{local}}, \mathbf{W}_{\text{context}}] = \text{Softmax}\Big(\text{Conv}_{1\times1\times1}\big(\mathbf{F}_{\text{local}} \parallel \mathbf{F}_{\text{context}}\big)\Big), \tag{7}$$

where $\parallel$ denotes channel-wise concatenation. The sum of the two weight maps at any spatial location (x, y, z) strictly equals 1. The adaptively integrated feature is then computed via an element-wise weighting summation:

$$\mathbf{F}_{\text{adapt}} = \mathbf{W}_{\text{local}} \odot \mathbf{F}_{\text{local}} + \mathbf{W}_{\text{context}} \odot \mathbf{F}_{\text{context}}. \tag{8}$$

Finally, the ASR module employs a residual connection to ensure stable optimization. The adaptively integrated feature is passed through a convolutional layer for channel alignment before being added back to the initial input:

$$\mathbf{F}_{\text{refined}} = \text{Conv}(\mathbf{F}_{\text{adapt}}) + \hat{\mathbf{F}}_{\text{fused}}. \tag{9}$$

where $\text{Conv}(\cdot)$ corresponds to the standard 3D convolution for channel projection. The output $\mathbf{F}_{\text{refined}}$ is subsequently forwarded to the task head.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our proposed VQSOP on the Semantic-OPV2V dataset introduced by CoHFF [26]. This dataset is an augmented version of the large-scale collaborative perception benchmark OPV2V [41], which is co-simulated by the OpenCDA [39] and CARLA [8] frameworks. In these simulated scenes, 2 to 7 connected vehicles exchange information, with each vehicle equipped with a 3D LiDAR sensor and four directional cameras. Because the original OPV2V dataset lacks semantic occupancy annotations, Semantic-OPV2V replays the original simulations to capture additional semantic LiDAR sweeps, thereby constructing reliable collaborative semantic occupancy supervision by aggregating the multi-agent ground truth.

Implementation Details. Following the experimental settings of CoHFF, the perception range is defined as $40 \times 40 \times 3.2$ m, which is discretized into a voxel grid of size $100\times100\times8$ with a voxel resolution of 0.4 m. For optimization, we employ the AdamW [20] optimizer with a weight decay of 0.01. The learning rate undergoes a linear warmup to 2×10^{-4} during the first 500 iterations, subsequently following a cosine annealing decay schedule. The entire framework is trained for 60 epochs with a batch size of 1 on a single NVIDIA RTX 4090 GPU. To supervise collaborative 3D occupancy prediction, the total loss consists of four terms: a voxel-wise cross-entropy loss for semantic classification, a scene-class affinity loss [2] to enforce structural consistency, a confidence loss for spatial mask learning, and the quantization loss.

Evaluation Metrics. Following previous works [3,18,26] on semantic occupancy prediction, we adopt mean Intersection-over-Union (mIoU) and class-wise

Table 1: 3D semantic occupancy prediction results on Semantic-OPV2V. Our proposed method achieves state-of-the-art performance under both single-agent and collaborative perception settings. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Single-Agent			Collaborative Perception		
	CoHFF [26]	GaussianFormer [15]	**Ours**	CoHFF [26]	GSFusion [3]	**Ours**
IoU	38.52	67.76	**70.40**	50.46	72.87	**73.79**
mIoU	24.85	29.20	**33.61**	34.16	37.44	**41.54**
Building	**21.04**	3.84	10.36	**25.72**	9.61	11.22
Fence	20.50	14.10	**30.16**	27.83	29.20	**33.19**
Terrain	43.93	**68.97**	48.26	48.30	**74.51**	63.19
Pole	**31.66**	5.94	12.25	**42.74**	12.19	20.16
Road	55.83	**79.37**	75.63	61.77	83.05	**87.06**
Side walk	17.31	**70.55**	57.72	39.62	**78.22**	67.40
Vegetation	14.49	12.54	**23.40**	20.59	20.43	**26.78**
Vehicles	**58.55**	49.25	52.37	63.28	60.49	**63.36**
Wall	33.30	30.79	**35.10**	**58.27**	36.45	36.51
Guard rail	1.54	15.01	**28.78**	1.94	32.50	**54.20**
Traffic signs	0.00	0.00	**13.36**	**16.33**	8.26	14.02
Bridge	0.00	0.00	**15.95**	3.53	4.35	**21.37**

Intersection-over-Union (IoU) as the primary evaluation metrics. Additionally, to assess performance in subsequent applications, we report the BEV 2D IoU for comparison with other baselines. Specifically, the predicted 3D voxel grids are projected onto the BEV plane along the height dimension to generate 2D semantic maps for evaluation.

4.2 Main Results

3D Semantic Occupancy Prediction. Table 1 presents the quantitative comparison of 3D semantic occupancy prediction on the Semantic-OPV2V dataset. Compared with CoHFF [26], which adopts a TPV representation, and the Vision-Only Gaussian Splatting framework proposed by Chen et al. [3] (hereafter referred to as GSFusion), our VQSOP achieves notable improvements across all metrics. Under the single-agent setting, our method reaches 70.40% in overall IoU and 33.61% in mIoU, significantly outperforming GaussianFormer [15] by 4.41% in mIoU. When extended to the collaborative perception setting, our VQSOP achieves consistent IoU improvements across all semantic categories, validating the necessity and effectiveness of multi-agent information sharing. In this setting, our method achieves the highest performance with 73.79% IoU and 41.54% mIoU, surpassing the second-best approach (GSFusion) by 4.10% in mIoU. Furthermore, our method achieves the best or second-best results across all semantic categories. Notably, class-wise analysis further reveals our method's strength in predicting small, thin, and geometrically complex objects. For example, under collaborative perception, our model achieves absolute IoU improvements of 21.70% on Guard rail and 17.02% on Bridge compared to the second-best results. These substantial gains suggest that the proposed ASR module plays a critical role in modeling fine-grained structures.

Table 2: Quantitative results on 2D BEV semantic segmentation. We report the 2D IoU for Vehicle, Road, and Others classes under two collaborative settings: 2 agents and up to 7 agents.

Approach	# Agents	Vehicle	Road	Others
CoBEVT [38]	2	46.13	52.41	-
CoHFF [26]	2	47.40	63.36	40.27
GSFusion [3]	2	70.25	82.69	79.37
Ours	2	**73.23**	**85.05**	**79.42**
CoBEVT [38]	Up to 7	60.40	63.00	-
CoHFF [26]	Up to 7	64.44	57.28	45.89
GSFusion [3]	Up to 7	75.30	84.96	80.19
Ours	Up to 7	**77.48**	**87.04**	**81.10**

BEV Segmentation. Table 2 presents the quantitative comparison of BEV segmentation under two collaborative settings. Our method consistently outperforms existing approaches in both scenarios. With 2 agents, we achieve 73.23% IoU on Vehicle and 85.05% on Road, surpassing the strongest baseline by 2.98% and 2.36%, respectively. When scaling up to 7 agents, the performance further improves to 77.48% on Vehicle and 87.04% on Road, maintaining clear margins over prior methods. These results indicate that the learned spatial representations remain robust and semantically consistent when applied to BEV segmentation.

Communication Volume. Table 3 compares the communication volume (CV) and the corresponding perception performance. As observed in the baseline methods, when GSFusion reduces the number of Gaussians from 25,600 to 6,400 to lower the CV from 1.07 MB to 0.27 MB, its mIoU drops from 37.44% to 36.02%, indicating a strict coupling between performance and bandwidth. In contrast, our method effectively breaks this bottleneck. It requires a CV of only 0.013 MB, which is 60× smaller than CoHFF (0.78 MB) and over 82× smaller than the high-resolution setting of GSFusion (1.07 MB). At this minimal bandwidth, our method achieves the highest performance with 73.79% IoU and 41.54% mIoU. This extreme communication efficiency demonstrates the effectiveness of our SAVQ module, which successfully isolates and quantizes only the essential foreground features, significantly reducing transmission costs.

4.3 Ablation Study

Component Analysis. To validate the effectiveness of our core modules, we conduct an ablation study as presented in Table 4. Without the SAVQ and ASR modules, the baseline achieves 72.48% IoU and 40.72% mIoU with a CV of 7.32 MB. When solely integrating the SAVQ module, the CV is drastically compressed from 7.32 MB to 0.013 MB. However, this extreme compression introduces a slight information loss, dropping the mIoU to 40.11%. Conversely, applying only the ASR module improves the baseline mIoU to 41.26% by refining the structural

Table 3: Comparison of communication volume and performance. CV denotes the Communication Volume per agent. ↓ indicates lower is better, and ↑ indicates higher is better. Our method achieves the lowest bandwidth requirement while maintaining the highest performance.

Method	CV (MB) ↓	IoU ↑	mIoU ↑
CoHFF [26]	0.78	50.46	34.16
GSFusion [3] (25600 Gaussians)	1.07	72.87	37.44
GSFusion [3] (6400 Gaussians)	0.27	72.42	36.02
Ours	**0.013**	**73.79**	**41.54**

Table 4: Ablation study on the core components of our proposed framework. We evaluate the individual and combined effects of the SAVQ and ASR modules on communication volume and perception accuracy.

SAVQ	ASR	CV (MB) ↓	mIoU ↑	IoU ↑
		7.32	40.72	72.48
✓		0.013	40.11	72.06
	✓	7.32	41.26	72.88
✓	✓	**0.013**	**41.54**	**73.79**

representations, though the CV remains at 7.32 MB. Notably, when both modules are enabled, the model achieves the best overall performance (73.79% IoU and 41.54% mIoU) under the minimal communication budget, demonstrating a strong synergistic effect between the two components.

Effect of Confidence Threshold. To investigate the impact of the confidence threshold τ within the Sparse-Aware selector of the SAVQ module, we conduct an ablation study as shown in Table 5. As τ increases from 0.6 to 0.8, the CV steadily decreases from 0.018 MB to 0.013 MB. Interestingly, the perception accuracy concurrently improves, peaking at 41.54% mIoU and 73.79% IoU. This trend suggests that a well-calibrated threshold effectively filters out task-irrelevant background noise, thereby refining the transmitted features and alleviating interference. However, when the threshold is excessively raised to 0.9, although the CV drops to its minimum of 0.012 MB, the performance suffers a noticeable decline (dropping to 41.03% mIoU and 72.24% IoU). This degradation indicates that an overly strict threshold inadvertently discards essential foreground structural details. Consequently, we set $\tau = 0.8$ as the default configuration to achieve the optimal balance between communication efficiency and perception accuracy.

Effectiveness of ASR Components. To validate the effectiveness of each ASR component, we conduct a fine-grained ablation study at the module level. As shown in Table 6, the first row corresponds to the setting without ASR. When both local and context branches are enabled but adaptive integration is disabled, we use fixed equal-weight fusion. The results show that both branches improve

Table 5: Ablation study on the confidence threshold τ in the SAVQ module. We evaluate the impact of different threshold values on both communication volume and accuracy.

Threshold	CV (MB) ↓	mIoU ↑	IoU ↑
0.6	0.018	41.47	73.57
0.7	0.015	41.52	73.66
0.8	0.013	**41.54**	**73.79**
0.9	**0.012**	41.03	72.24

Table 6: Ablation study on the ASR module with SAVQ enabled. We evaluate the individual contributions of the local branch, context branch, and adaptive integration mechanism.

Local	Context	Adaptive	mIoU ↑	IoU ↑
			40.11	72.06
✓			40.23	72.56
	✓		40.85	72.68
✓	✓		41.03	73.14
✓	✓	✓	**41.54**	**73.79**

over the setting without ASR, and the context branch achieves higher mIoU, indicating the importance of broader semantic context for feature refinement. The last two rows show the benefit of adaptive integration in better leveraging the complementarity of the two branches.

Visualization. To provide a more intuitive understanding of the impact of the ASR module, we present qualitative comparisons in Fig. 5. The red boxes highlight representative regions containing fine-grained structures and boundary-sensitive areas. For VQSOP without ASR, the model produces incomplete predictions with noticeable holes in road regions and occasionally misses small objects and thin structures, such as poles and fences. These discrepancies are more evident in distant regions and geometrically complex areas. In contrast, VQSOP with ASR aligns more closely with the ground truth. The full model better reconstructs fine-grained geometric details and preserves structural continuity, resulting in more coherent scene layouts. These observations highlight the important role of the ASR module in refining local spatial context and improving the modeling of complex scene geometries.

5 Conclusion

In this paper, we proposed VQSOP, a bandwidth-efficient framework for collaborative 3D semantic occupancy prediction. The proposed approach encodes high-dimensional 3D features into compact discrete representations, significantly reducing transmission overhead while preserving essential geometric semantics.

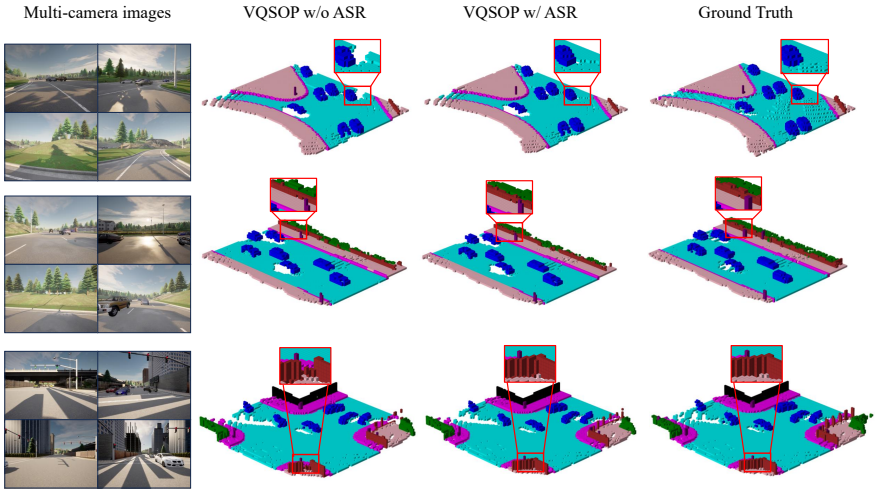

Fig. 5: Qualitative results of 3D semantic occupancy prediction. From left to right: the input multi-camera images, predictions of VQSOP w/o ASR, our full model VQSOP w/ ASR, and the Ground Truth. The red zoomed-in regions highlight that our VQSOP equipped with ASR successfully recovers fine-grained geometric details.

Through spatial refinement after collaborative fusion, the framework further improves structural completeness and semantic consistency. Extensive experiments on the Semantic-OPV2V dataset demonstrate that VQSOP achieves state-of-the-art performance with lower communication volume, providing an effective and practical solution for communication-efficient collaborative 3D semantic occupancy prediction.

References

1. Arnold, E., Dianati, M., de Temple, R., Fallah, S.: Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. IEEE Transactions on Intelligent Transportation Systems **23**(3), 1852–1864 (2022)
2. Cao, A.Q., De Charette, R.: MonoScene: Monocular 3D semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022)
3. Chen, C., Huang, H., Bagchi, S.: Vision-only gaussian splatting for collaborative semantic occupancy prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2796–2804 (2026)
4. Chen, G., Zhang, C., Zhao, X.: WhisperNet: A scalable solution for bandwidth-efficient collaboration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 32154–32163 (2026)
5. Chen, Q., Tang, S., Yang, Q., Fu, S.: Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds. In: 2019 IEEE 39th International Conference on distributed computing systems (ICDCS). pp. 514–524. IEEE (2019)

6. Cheng, R., Agia, C., Ren, Y., Li, X., Bingbing, L.: S3CNet: A sparse semantic scene completion network for lidar point clouds. In: Conference on Robot Learning. pp. 2148–2161. PMLR (2021)
7. Cui, J., Qiu, H., Chen, D., Stone, P., Zhu, Y.: Coopernaut: End-to-end driving with cooperative perception for networked vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17252–17262 (2022)
8. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
9. Duan, Z., Dang, C., Hu, X., An, P., Ding, J., Zhan, J., Xu, Y., Ma, J.: SDGOCC: Semantic and depth-guided bird's-eye view transformation for 3D multimodal occupancy prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6751–6760 (2025)
10. Gao, X., Zhang, X., Lu, Y., Huang, Y., Yang, L., Xiong, Y., Liu, P.: A survey of collaborative perception in intelligent vehicles at intersections. IEEE Transactions on Intelligent Vehicles (2024)
11. Hu, Y., Fang, S., Lei, Z., Zhong, Y., Chen, S.: Where2comm: Communication-efficient collaborative perception via spatial confidence maps. Advances in neural information processing systems **35**, 4874–4886 (2022)
12. Hu, Y., Lu, Y., Xu, R., Xie, W., Chen, S., Wang, Y.: Collaboration helps camera overtake lidar in 3D detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9243–9252 (2023)
13. Hu, Y., Peng, J., Liu, S., Ge, J., Liu, S., Chen, S.: Communication-efficient collaborative perception via information filling with codebook. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15481–15490 (2024)
14. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3D semantic occupancy prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9223–9232 (2023)
15. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: GaussianFormer: Scene as gaussians for vision-based 3D semantic occupancy prediction. In: European Conference on Computer Vision. pp. 376–393. Springer (2024)
16. Li, J., He, X., Zhou, C., Cheng, X., Wen, Y., Zhang, D.: ViewFormer: Exploring spatiotemporal modeling for multi-view 3D occupancy perception via view-guided transformers. In: European Conference on Computer Vision. pp. 90–106. Springer (2024)
17. Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., Zhang, W.: Learning distilled collaboration graph for multi-agent perception. Advances in Neural Information Processing Systems **34**, 29541–29552 (2021)
18. Li, Y., Yu, Z., Choy, C., Xiao, C., Alvarez, J.M., Fidler, S., Feng, C., Anandkumar, A.: VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9087–9098 (2023)
19. Liu, H., Chen, Y., Wang, H., Yang, Z., Li, T., Zeng, J., Chen, L., Li, H., Wang, L.: Fully sparse 3D occupancy prediction. In: European Conference on Computer Vision. pp. 54–71. Springer (2024)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Mei, J., Yang, Y., Wang, M., Zhu, J., Ra, J., Ma, Y., Li, L., Liu, Y.: Camera-based 3D semantic scene completion with sparse guidance network. IEEE Transactions on Image Processing **33**, 5468–5481 (2024)

22. Omeiza, D., Webb, H., Jirotka, M., Kunze, L.: Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems **23**(8), 10142–10162 (2022)
23. Pradeep, A., Bakoev, M., Akhroljonova, N.: A reliability analysis of self-driving vehicles: evaluating the safety and performance of autonomous driving systems. In: 2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). pp. 1–5. IEEE (2023)
24. Qiao, D., Zulkernine, F., Anand, A.: CoBEVFusion cooperative perception with lidar-camera bird's eye view fusion. In: 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). pp. 389–396. IEEE (2024)
25. Shi, S., Cui, J., Jiang, Z., Yan, Z., Xing, G., Niu, J., Ouyang, Z.: VIPS: Real-time perception fusion for infrastructure-assisted autonomous driving. In: Proceedings of the 28th annual international conference on mobile computing and networking. pp. 133–146 (2022)
26. Song, R., Liang, C., Cao, H., Yan, Z., Zimmer, W., Gross, M., Festag, A., Knoll, A.: Collaborative semantic occupancy prediction with hybrid feature fusion in connected automated vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17996–18006 (2024)
27. Su, S., Han, S., Li, Y., Zhang, Z., Feng, C., Ding, C., Miao, F.: Collaborative multi-object tracking with conformal uncertainty propagation. IEEE Robotics and Automation Letters **9**(4), 3323–3330 (2024)
28. Tan, J., Lyu, F., Li, L., Hu, F., Feng, T., Xu, F., Zhang, Z., Yao, R., Wang, L.: Dynamic V2X perception from road-to-vehicle vision. IEEE Transactions on Intelligent Vehicles (2024)
29. Tang, P., Wang, Z., Wang, G., Zheng, J., Ren, X., Feng, B., Ma, C.: SparseOcc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15035–15044 (2024)
30. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
31. Wang, B., Zhang, L., Wang, Z., Zhao, Y., Zhou, T.: CORE: Cooperative reconstruction for multi-agent perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8710–8720 (2023)
32. Wang, J., Liu, Z., Meng, Q., Yan, L., Wang, K., Yang, J., Liu, W., Hou, Q., Cheng, M.M.: OPUS: occupancy prediction using a sparse set. Advances in Neural Information Processing Systems **37**, 119861–119885 (2024)
33. Wang, T., Kim, S., Wenxuan, J., Xie, E., Ge, C., Chen, J., Li, Z., Luo, P.: DeepAccident: A motion and accident prediction benchmark for V2X autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5599–5606 (2024)
34. Wang, T.H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., Urtasun, R.: V2VNet: Vehicle-to-vehicle communication for joint perception and prediction. In: European conference on computer vision. pp. 605–621. Springer (2020)
35. Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., Lu, J.: SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21729–21740 (2023)
36. Wu, H., Lin, P., Javanmardi, E., Bao, N., Qian, B., Si, H., Tsukada, M.: A synthetic benchmark for collaborative 3D semantic occupancy prediction in V2X autonomous driving. arXiv preprint arXiv:2506.17004 (2025)

37. Xiang, H., Xu, R., Ma, J.: HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 284–295 (2023)
38. Xu, R., Tu, Z., Xiang, H., Shao, W., Zhou, B., Ma, J.: CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers. arXiv preprint arXiv:2207.02202 (2022)
39. Xu, R., Xiang, H., Han, X., Xia, X., Meng, Z., Chen, C.J., Correa-Jullian, C., Ma, J.: The opencda open-source ecosystem for cooperative driving automation research. IEEE Transactions on Intelligent Vehicles **8**(4), 2698–2711 (2023)
40. Xu, R., Xiang, H., Tu, Z., Xia, X., Yang, M.H., Ma, J.: V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In: European conference on computer vision. pp. 107–124. Springer (2022)
41. Xu, R., Xiang, H., Xia, X., Han, X., Li, J., Ma, J.: OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2583–2589. IEEE (2022)
42. Yan, X., Gao, J., Li, J., Zhang, R., Li, Z., Huang, R., Cui, S.: Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 3101–3109 (2021)
43. Yang, D., Huang, S., Xu, Z., Li, Z., Wang, S., Li, M., Wang, Y., Liu, Y., Yang, K., Chen, Z., et al.: AIDE: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20459–20470 (2023)
44. Yang, K., Yang, D., Zhang, J., Li, M., Liu, Y., Liu, J., Wang, H., Sun, P., Song, L.: Spatio-temporal domain awareness for multi-agent collaborative perception. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 23383–23392 (2023)
45. Zhang, Y., Zhu, Z., Du, D.: OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9433–9443 (2023)
46. Zuo, S., Zheng, W., Huang, Y., Zhou, J., Lu, J.: PointOcc: Cylindrical tri-perspective view for point-based 3D semantic occupancy prediction. arXiv preprint arXiv:2308.16896 (2023)