

Communication-Efficient Multi-Vehicle Collaborative Semantic Segmentation via Sparse 3D Gaussian Sharing

Tianyu Hong¹ Xiaobo Zhou^{1,2*} Wenkai Hu¹ Qi Xie¹ Zhihui Ke¹ Tie Qiu^{1,2}

¹Tianjin University, China ²Qinghai Minzu University, China

{tianyuhong, xiaobo.zhou, wenkai.hu, xie.qi, kezhihui, qiutie}@tju.edu.cn

Abstract

Collaborative perception is considered a promising approach to address the inherent limitations of single-vehicle systems by sharing data among vehicles, thereby enhancing performance in perception tasks such as bird’s-eye view (BEV) semantic segmentation. However, existing methods share the entire dense, scene-level BEV feature, which contains significant redundancy and lacks height information, ultimately leading to unavoidable bandwidth waste and performance degradation. To address these challenges, we present GSCOOP, the first collaborative semantic segmentation framework that leverages sparse, object-centric 3D Gaussians to fundamentally overcome communication bottlenecks. By representing scenes with compact Gaussians that preserve complete spatial information, GSCOOP achieves both high perception accuracy and communication efficiency. To further optimize transmission, we introduce the Priority-Based Gaussian Selection (PGS) module to adaptively select critical Gaussians and a Semantic Gaussian Compression (SGC) module to compress Gaussian attributes with minimal overhead. Extensive experiments on OPV2V and V2X-Seq demonstrate that GSCOOP achieves state-of-the-art performance, even with more than $500\times$ lower communication volume. The code link is <https://github.com/SHEVIP/GSCOOP>.

1. Introduction

Bird’s Eye View (BEV) semantic segmentation plays a crucial role in autonomous driving perception systems [22, 23, 25], as it offers more comprehensive scene information compared to object detection [20]. Although the perception capabilities of single-vehicle systems have been significantly improved, they are still limited by inherent challenges such as long-range detection and occlusion [4, 17, 35, 36, 45]. Vehicle-to-Vehicle (V2V) communication technology, which enables the exchange of perception information between multi-vehicle, has experienced rapid development in recent

*Corresponding author: xiaobo.zhou@tju.edu.cn

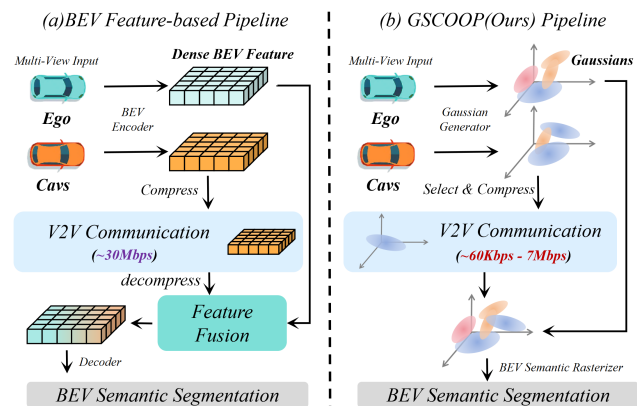


Figure 1. Comparison of pipeline in collaborative semantic segmentation. (a) BEV Feature-based pipeline rely on bandwidth-inefficient dense BEV features and suffer from performance degradation due to lack of height information. (b) GSCOOP (Ours) pipeline generates discrete and object-centric 3D Gaussians representations, achieving accurate scene representation with minimal bandwidth.

years. Subsequently, collaborative BEV semantic segmentation based on cameras has gained increasing attention [13, 28, 30, 31, 38, 41, 42, 44, 46], which provides a vision-based solution for the deployment of collaborative systems in real-world scenarios without costly LiDAR sensors.

Existing methods [7, 28, 30–32, 38] (see Fig.1(a)) represent scenes using BEV feature map extracted from multi-view images. This BEV feature map are compressed by collaborative CAVs for transmission and then decompressed, aligned, and fused with local features by the ego vehicle to predict the BEV semantic segmentation map. However, the 2D BEV feature maps inevitably suffer from the loss of height information [10], which limits the effectiveness of feature fusion and degrades segmentation accuracy. Moreover, transmitting the entire BEV feature map, typically requiring around 30 Mbps, not only exceeds the practical C-V2X bandwidth limit of 10 Mbps [21] but also introduces substantial redundancy. Even more critically, extracting and transmitting only the most relevant regions is inherently

challenging, as BEV feature maps densely encode the entire scene in a continuous manner, lacking explicit sparsity to facilitate efficient selection. These limitations motivate us to explore a novel framework in collaborative perception that enables efficient transmission while preserving critical scene information.

In this paper, we introduce *GSCOOP*, the first 3D Gaussian-based collaborative semantic segmentation framework leveraging sparse Gaussian sharing, as illustrated in Fig. 1(b). Unlike continuous BEV feature maps, we use Gaussians with explicit spatial and semantic attributes to discretely and accurately represent the objects in the scene, allowing selective transmission of only the most informative sparse Gaussians. Specifically, the Semantic Gaussian Generator (SGG) efficiently extracts compact 3D Gaussians from multi-view images, providing accurate spatial and semantic representations of the scene. To minimize communication overhead, the Priority-Based Gaussian Selection (PGS) module adaptively selects sparse Gaussians based on their priority, while the Semantic Gaussian Compression (SGC) module compactly encodes Gaussian attributes, reducing bandwidth usage to just 62 *Kbps*-7 *Mbps*. Ego vehicle obtains a more complete scene representation through simple spatial alignment, effectively avoiding the heavy computation of BEV feature fusion networks. Finally, the ego vehicle directly renders the BEV semantic segmentation from the aggregated Gaussians using the proposed BEV Semantic Gaussian Rasterizer. In addition, we are the first to evaluate collaborative semantic segmentation on the real-world dataset V2X-Seq [43], rather than only on the commonly used simulation dataset OPV2V [37]. Overall, the main contributions of this paper can be summarized as follows:

- We propose *GSCOOP*, the first collaborative semantic segmentation framework based on 3D Gaussian representations. A Semantic Gaussian Generator (SGG) efficiently constructs these representations, while a BEV Semantic Rasterizer directly renders segmentation maps via orthogonal projection.
- We develop a Priority-Based Gaussian Selection (PGS) module to adaptively select and transmit the most informative Gaussians under bandwidth constraints, together with a Semantic Gaussian Compression (SGC) module that applies a learnable codebook for further communication reduction.
- Extensive experiments show that *GSCOOP* achieves state-of-the-art performance, surpassing leading multi-agent fusion methods by up to 6.30% and 4.01% on OPV2V and V2X-Seq, respectively. Notably, our method reduces communication volume by nearly 500 times while maintaining comparable performance.

2. Related Work and Background

2.1. Cooperative Perception

Collaborative perception enables CAVs to enhance their perception by sharing information within a multi-vehicle system [1, 11, 12, 28, 31, 38, 39]. Most existing approaches operate on BEV representations, aiming to improve semantic segmentation and adapt to diverse driving scenarios through specialized designs. CoBEVT [38] introduces a Transformer-based FAX module for effective temporal and spatial fusion. CORE [31] leverages LiDAR data with spatial-channel compression and attention-aware collaboration to reconstruct scenes efficiently. CoBEVFusion [28] explores multi-modal fusion through a DWCA module, combining camera and LiDAR features via CNN-based inter-agent interaction. DI-V2X [16] addresses domain and spatial heterogeneity with adaptive attention mechanisms. While these works achieved notable progress, challenges such as missing height information and bandwidth inefficiency remain key limitations in collaborative semantic segmentation, which motivates our Gaussian-based approach.

2.2. 3D Gaussian Splatting

Recent progress in 3D Gaussian Splatting(3D-GS) [15] demonstrates its effectiveness in radiance field rendering, achieving high-quality results with real-time efficiency. By optimizing continuous Gaussian parameters through differentiable rendering, 3D-GS provides compact and expressive scene representations, outperforming dense voxel grids by allocating Gaussians adaptively to capture both coarse structures and fine details. Building on its strong representational capacity, 3D-GS has been extended to autonomous driving scenarios. Several studies [8, 26, 29, 34] incorporate semantics into Gaussians, while others [2, 6, 14, 40, 47] explore Gaussians for perception tasks such as occupancy prediction. GaussianFormer[14] reduces memory consumption by introducing Gaussian representations to the occupancy task. GaussianBeV[2] predicts per-pixel Gaussian distributions and renders semantic features. GaussianPretrain[40] leverages 3D Gaussian anchors to learn structural priors, improving pre-training across downstream tasks. In contrast to these single-vehicle methods, *GSCOOP* is the first to explore Gaussians in a collaborative setting. It introduces a fundamentally different voxel-based Gaussian generator, incorporates tailored selection and compression mechanisms, and provides an analysis of how different Gaussian attributes affect the accuracy–bandwidth trade-off.

3. Methodology

3.1. Overview

The overall architecture of *GSCOOP* is illustrated in Fig. 2, which consists of four main components: Semantic Gaus-

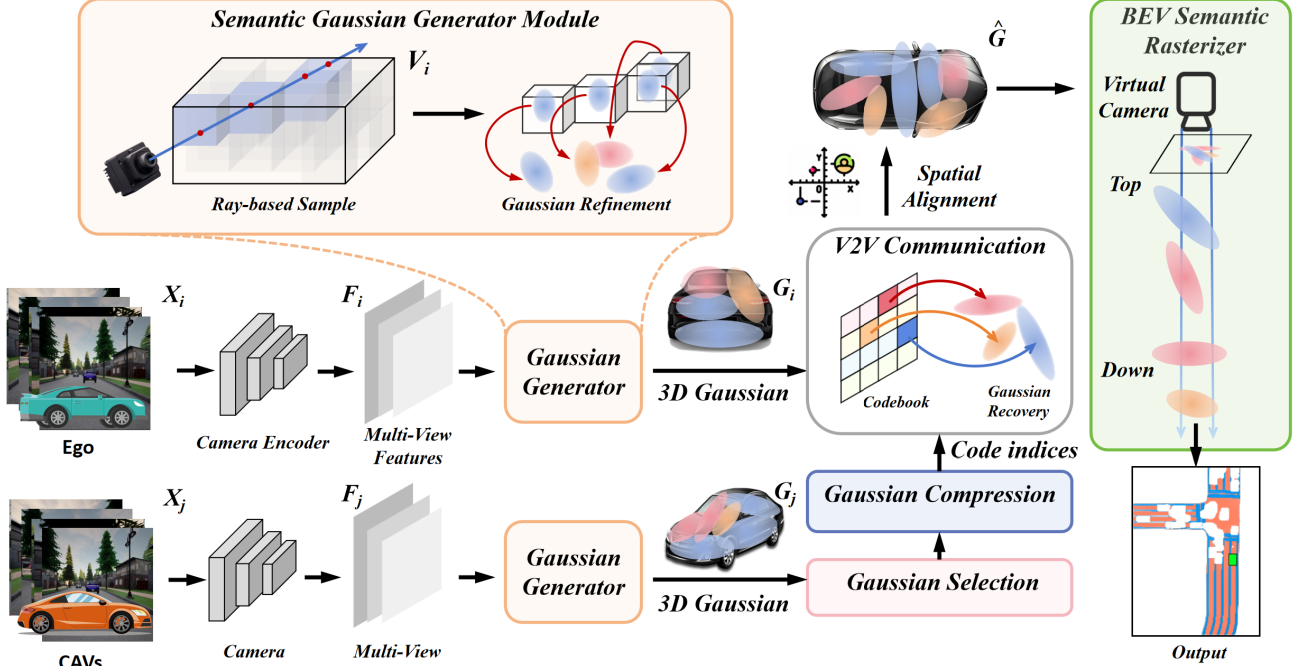


Figure 2. Overview of the proposed GSCOOP framework. Ego and collaborative CAVs extract multi-view features from images, which are transformed into semantic 3D Gaussians via the **Semantic Gaussian Generator** with ray-based sampling and refinement. Collaborative CAVs employ **Priority-Based Gaussian Selection** and **Semantic Gaussian Compression** scheme to transmit sparse Gaussian representation via V2V communication. The ego vehicle recovers the Gaussians from the received information, spatially aligns them with local Gaussians, and renders them into a BEV semantic map via orthogonal Gaussian splatting.

sian Generator module, Priority-Based Gaussian Selection, Semantic Gaussian Compression module, and Bev Semantic Rasterizer. These components work together to balance accuracy and bandwidth, ensuring scalable collaboration, with each component introduced in detail below.

3.2. Camera encoder and Semantic Gaussian Generator

Suppose there are M agents in the collaboration scenario, with the original multi-view images X_i of the i th agent as input. The camera encoder is used for extracting the multi-view features $F_i \in \mathbb{R}^{K \times H_F \times W_F \times C_F}$. Here, H_F and W_F denote the height and width of the feature maps, where K represents the number of cameras equipped on the vehicle, and C_F is the number of feature channels.

To efficiently capture both geometric and semantic details, we adopt the 'lift' process from LSS [27] to project multi-view image features into a 3D voxel space. Rather than relying on explicit depth estimation, we directly sample image features along viewing rays, thereby obtaining a direct voxel feature representation $V_i \in \mathbb{R}^{X \times Y \times Z \times C_F}$. Here X, Y, Z denote the dimensions along the $x, y,$ and z axes, respectively. Each voxel center serves as the initialization point for a Gaussian distribution g . Each Gaussian is represented by a vector in the form of $(\mu \in \mathbb{R}^3, s \in \mathbb{R}^3, r \in \mathbb{R}^4, c \in \mathbb{R}^{|C|}, \alpha \in \mathbb{R}^4)$ and μ, s, r, c, α, C denote the position, scale, rotation, se-

mantic logits, opacity and the number of semantic category. A Gaussian point in 3D space is defined as:

$$G(X) = \exp\left(-\frac{1}{2}(X - \mu)^\top \Sigma^{-1}(X - \mu)\right). \quad (1)$$

where Σ is the 3D covariance matrix. To achieve efficient optimization and flexible adjustment of each Gaussian's rotation and scaling, we adopt the approach of prior work [15] by decomposing the covariance matrix Σ into a rotation matrix R and a scaling matrix S . It is defined as:

$$\Sigma = RSS^\top R^\top, \quad S = \text{diag}(s), \quad R = \text{q2r}(r), \quad (2)$$

where $\text{diag}(\cdot)$ and $\text{q2r}(\cdot)$ represent the function that generate a diagonal matrix from a vector and convert a quaternion into a rotation matrix, respectively.

Considering the diverse attributes of Gaussian representations, originally designed for high-quality image rendering, we aim to identify the most effective Gaussian attributes for improving segmentation while minimizing bandwidth. Our experiments results as shown in Table.1, indicate that scale and rotation contribute minimally, likely due to their weak supervision in segmentation tasks. Meanwhile, we found redundancy between the meaning of the empty class in the semantics logits and the opacity attribute.

Table 1. Exploratory experiments the contribution of Gaussian attributes on the OPV2V dataset.

Pos.	Sca.	Rota.	Veh./Road/Lane \uparrow	Vol.(KB) \downarrow
			52.38/61.04/40.32	1.22
✓			57.21/64.77/46.71	15.87
✓	✓		57.59/64.12/47.27	30.52
✓		✓	57.39/64.39/47.10	35.40
✓	✓	✓	58.06/64.27/47.72	50.05

Based on these insights, our approach refines the Gaussian parameters by using a 3D convolutional neural network to directly decode the 3D voxel features. We predict the semantic logits $c = [c_1, c_2, \dots, c_C]$ and the position offsets $\Delta\mu = [\Delta X, \Delta Y, \Delta Z]$ between the refined Gaussian and the center of the voxel. Position offsets correct the limitations of fixed voxelization by enhancing spatial flexibility, while the semantic logits c_j reflects the probability that a Gaussian belongs to a specific category j . Then, the opacity α is computed by normalizing the logits of the empty category in c . Meanwhile, we supervise the spatial position of each category by regularizing the Gaussian distribution of the position and its semantic labels. Let $G = \{g^{(i)}\}_{i=1}^N$ denote the set of all Gaussians, where N denotes the total number of Gaussians. Each Gaussian $g^{(i)}$ with height $Z^{(i)}$ and semantic label $\hat{c}^{(i)} = \arg \max_j c_j^{(i)}$, the loss for each gaussian is defined as

$$\ell(G^{(i)}) = \frac{(Z^{(i)} - \mu_{z, \hat{c}^{(i)}})^2}{2\sigma_{z, \hat{c}^{(i)}}^2}, \quad (3)$$

where $\mu_{z, \hat{c}^{(i)}}$ and $\sigma_{z, \hat{c}^{(i)}}$ are predefined height priors. The total regularization loss is:

$$L_h = \frac{1}{N} \sum_{i=1}^N \ell(G^{(i)}). \quad (4)$$

3.3. Priority-Based Gaussian Selection Module

The core idea of the **Priority-Based Gaussian Selection** module is to select the most critical Gaussians for scene representation to minimize bandwidth consumption. We assess the priority of each Gaussian in two ways as shown in Fig. 3(a): The first is based on its impact on the final projection, where higher and more opaque objects contribute more significantly by providing richer information. Second, in collaborative settings, multiple vehicles often need to reach a consensus on ambiguously located objects for accurate perception; thus, Gaussians with uncertain semantic information may contain unconfirmed yet critical details. Additionally, even in scenarios with ample bandwidth, it is unnecessary to fully utilize the capacity, so we fine-tune the number of Gaussians transmitted based on their priority scores.

Relevance-Projection Analysis (RPA). This module evaluates the significance of each Gaussian $g^{(i)} \in G$ by considering its height and opacity, two critical factors in the

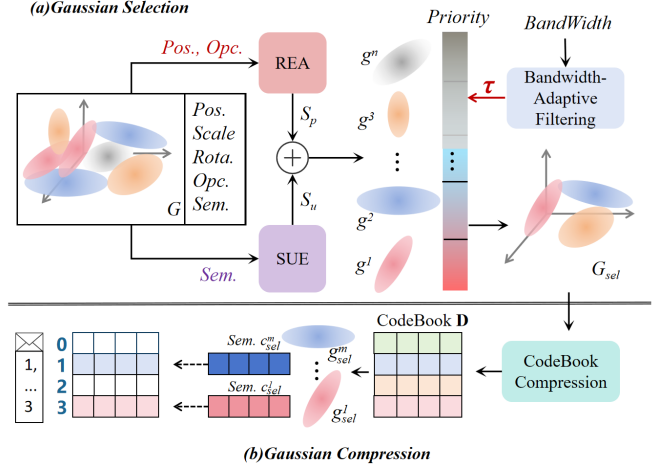


Figure 3. Process of Gaussian Selection in 3.3 and Semantic Compression in 3.4. Here **REA** stands for Relevance-Projection Analysis and **SUE** stands for Semantic Uncertainty Evaluation.

splatting process. The priority score of $g^{(i)}$ is computed as:

$$S_p^{(i)} = \text{RPA}(g^{(i)}) = \beta_h \cdot \sigma(\gamma_h \cdot Z^{(i)}) + \beta_o \cdot \alpha^{(i)}, \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid function to normalize the influence of height, preventing dominance from extreme values. The parameter γ_h controls the steepness of the sigmoid curve, while β_h and β_o are weighting factors balancing the contributions of height and opacity. Higher $S_p^{(i)}$ values indicate Gaussians with greater significance in the splatting process, prioritizing them for efficient transmission.

Semantic Uncertainty Evaluation (SUE). For Gaussian $g^{(i)}$, with semantic classification probability vector $\mathbf{c}^{(i)}$, the uncertainty score is calculated as:

$$S_u^{(i)} = \text{SUE}(g^{(i)}) = -\beta_u \sum_{j=1}^C c_j^{(i)} \log c_j^{(i)}, \quad (6)$$

where β_u is a weighting factor to adjust the influence of uncertainty. This entropy-based metric quantifies the semantic ambiguity of $G^{(i)}$. Gaussians with higher $S_u^{(i)}$ are prioritized, as their uncertain yet potentially critical semantic information can enhance collaborative perception in multi-vehicle scenarios.

Bandwidth-Adaptive Filtering module dynamically selects and transmits the most critical Gaussians. Instead of fixed thresholds, the module computes a transmission threshold τ_0 based on the available bandwidth B_0 and the priority score set $P(G) = \{P(g^{(i)})\}_{i=1}^N$, where $P(g^{(i)}) = \sigma(S_p^{(i)} + S_u^{(i)})$. We fine-tune the threshold as follows:

$$\tau = f(B, P(G)), B = B_0 \left(1 + \beta_d \frac{(\tau_0 - \bar{I}(P(G)))}{\sigma(P(G))} \right), \quad (7)$$

where $f(\cdot)$ is a function that computes the minimum threshold corresponding to the bandwidth. Here, $\bar{I}(\cdot)$ and $\sigma(\cdot)$ represent the mean and standard deviation of the priority scores, respectively. And the β_d is a weighting factor. This ensures that the selection process adapts to both network conditions and the distribution of importance scores. Finally, the transmitted subset is determined by:

$$G_{sel} = \{g^{(i)} \in G \mid P(g^{(i)}) \geq \tau\}. \quad (8)$$

3.4. Semantic Gaussian Compression Module

To efficiently compress high-dimensional Gaussian semantic data, we introduce a **Gaussian semantic compression** scheme using learnable codebooks. Inspired by [24, 34], we utilize learnable parameters to adapt the codebook for better performance. Each data point is mapped to the closest codebook vector, enabling transmission of one-byte indexes instead of the full Gaussian semantic attribute composed of floating-point numbers.

Codebook Learning. Given a set of Gaussians $G_{sel} = \{g_{sel}^{(i)}\}_{i=1}^M$ after selection, we extract their semantic logits $\mathbf{C}_{sel} = \{\mathbf{c}_{sel}^{(i)}\}_{i=1}^N$, where $\mathbf{c}_{sel}^{(i)} \in \mathbb{R}^C$ denotes the semantic logits of the i -th Gaussian. We define the codebook as $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_{n_D}] \in \mathbb{R}^{C \times n_D}$, where n_D is the number of codebook vectors. For each $\mathbf{c}^{(i)}$, the nearest codeword is assigned by:

$$q^{(i)} = \arg \min_{\ell \in \{1, \dots, n_D\}} \|\mathbf{c}_{sel}^{(i)} - \mathbf{d}_\ell\|_2^2. \quad (9)$$

The quantized semantic representation is $\hat{\mathbf{c}}^{(i)} = \mathbf{d}_{q^{(i)}}$. The codebook \mathbf{D} is optimized to minimize the reconstruction error over all Gaussians:

$$\mathcal{L}_{rec} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{c}_{sel}^{(i)} - \hat{\mathbf{c}}^{(i)}\|_2^2. \quad (10)$$

Through backpropagation, \mathbf{D} is iteratively refined to accurately capture semantic structures within Gaussians.

Codebook Compression. After training, the codebook \mathbf{D} is fixed, and each Gaussian's semantic logits \mathbf{c} are compressed by assigning the nearest codeword index. This reduces the transmission from high-dimensional logits $\mathbf{c} \in \mathbb{R}^C$ to a compact index $q \in \{1, \dots, n_D\}$, significantly improving communication efficiency while preserving essential semantics.

3.5. BeV Semantic Rasterizer and Loss Function

The ego vehicle decodes semantic information from the indices q , merges the received Gaussians with local Gaussians into the final set \hat{G}_i after spatial alignment.

The BeV Semantic Rasterizer is used to obtain the BEV semantic segmentation map $S \in \mathbb{R}^{H_B \times W_B}$ from \hat{G} . To

adapt Gaussian Splatting [15] for BEV perception, we modify the rendering pipeline by replacing perspective projection with orthogonal projection, preserving object shapes and sizes in the BEV view. Additionally, while the original process projects Gaussian colors to generate RGB images, we project semantic attributes to produce segmentation maps, following offline semantic reconstruction practices [29, 34]. Finally, the pixel semantic segmentation is rendered from N Gaussians contributing to this pixel in sorted order with the blending equation:

$$S = \arg \max \left(\text{Softmax} \left(\sum_{i=1}^N c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \right) \right), \quad (11)$$

where $\prod_{j=1}^{i-1} (1 - \alpha_j)$ is the transmittance.

To train our module, we use the standard semantic loss function, and the total loss L_{sem} is defined as follows:

$$L_{sem} = \lambda_d L_d + \lambda_s L_s + \lambda_h L_h, \quad (12)$$

where λ_d , λ_s and λ_h are weighting factors. L_d and L_s correspond to the binary cross-entropy loss between the predicted values S and the ground truth for dynamic and static objects, respectively.

4. Experiment

4.1. Datasets

To evaluate the performance of our proposed GSCOOP, we conducted extensive experiments on both real-world and simulated datasets: V2X-Seq [43] and OPV2V [37]. The V2X-Seq dataset is a real-world sequential V2X dataset that includes over 15,000 frames from 95 real-world scenarios, each collaboration scenario includes one vehicle and one infrastructure equipped with one camera. The OPV2V dataset is a simulated dataset collected using the OpenCDA [37] and CARLA [5] framework with 2 to 7 vehicles equipped with one LiDAR sensor and four cameras. We used vector maps and object detection ground truth to generate semantic segmentation ground truth in V2X-Seq. Our primary experiments utilized the camera sensors within both datasets, employing Intersection over Union (IoU) between the predicted maps and ground truth maps as the performance metric. Both datasets were evaluated using the same setup in [38]. Ground truth categories in both datasets included dynamic vehicles, drivable roads, and lane lines, enabling a consistent evaluation framework across the real-world and simulated environments.

4.2. Implementation details

Experimental Setup. We configured all autonomous vehicles with a communication range of 70 meters, following the guidelines in [33]. We limit the evaluation range to

Table 2. Performance comparison on the task of BeV Semantic Segmentation on OPV2V and V2X-Seq datasets.

Model	OPV2V IOU(%)				V2X-Seq IOU(%)			
	Vehicle	Road	Lane	Avg	Vehicle	Road	Lane	Avg
No Fusion	26.50	40.30	27.69	31.50	28.12	66.34	23.77	39.41
F-Cooper[3]	44.35	49.99	32.45	42.26	42.56	79.88	34.63	52.36
AttFuse[37]	32.94	45.23	30.82	36.33	43.23	79.68	35.40	52.77
CoBEVT[38]	52.95	55.12	41.62	49.90	45.24	81.89	40.14	55.76
DI-V2X[16]	48.67	53.42	38.83	46.97	44.46	81.83	37.82	54.70
ERMVP[46]	48.33	56.78	41.80	48.97	43.28	83.76	42.40	56.48
GSCOOP(Ours)	57.21	64.67	46.71	56.20	47.61	88.41	45.44	60.49

$[-50, 50]m \times [-50, 50]m \times [-3, 1]m$ follow [38], the ground truth map is 256×256 with a 39 cm map resolution. Our model is trained end-to-end on collaborative scenario data for 70 epochs, without separately training static and dynamic models, with a batch size of 8 per GPU. We jointly trained all models for fair comparison without separating static and dynamic models. All models are trained on 2 NVIDIA RTX A6000 GPUs for uniformity and comparability of results.

Training details. For both datasets, we employed ResNet34 [9] as the camera feature encoder. The predefined scene Gaussian representation consisted of $100 * 100 * 4$ Gaussians. For Gaussian selection, we first use an opacity threshold of 0.65 to remove useless Gaussian. The size of the codebook is set to 64 and it is trained separately for 20 rounds using only \mathcal{L}_{rec} , with other parameters frozen. As in [38], λ_d, λ_s were set to 2,1 and we set λ_h to 5×10^{-7} . The Adam optimizer [18] and cosine annealing learning rate scheduler [19] with a learning rate of 3×10^{-4} were used for training.

4.3. Main results

Task Performance Evaluation. Table.2 shows the BeV segmentation performance comparison results on the OPV2V and V2X-Seq datasets. We used the single-agent perception (No Fusion) as the baseline. Meanwhile, the existing state-of-the-art multi-agent perception algorithms are fully considered: F-Cooper [3], AttFuse[37], CoBevt [38] and DI-V2X[16], ERMVP[46]. The proposed GSCOOP outperforms all other SOTA models in both simulation and real scenarios, demonstrating its ability to accurately capture detailed spatial features. In particular, the SOTA performance is improved by 6.30% and 4.01% on the OPV2V and V2X-Seq datasets separately in average IOU. Significantly improved performance on 'Vehicle' and 'Lane' categories that require more precise representational capabilities demonstrates the advancement of the proposed framework.

Comparison of Communication Volume. Evaluating performance across different bandwidths is essential for real-world perception applications. To this purpose, Fig. 4

compares the perception performance of our GSCOOP with CoBEVT[38] under varying communication volumes. Note that we used the 64×64 intermediate features from the official CoBeVT implementation for better contrast, while other methods follow the official settings. We observe that: (i) GSCOOP maintains strong performance across all communication volumes, even outperforming some methods under extremely low communication volumes.(ii) GSCOOP outperforms CoBevt at every communication volume, with only a slight decrease in performance even under limited bandwidth. This is due to the efficiency of our Gaussian-based semantic filtering and adaptive bandwidth selection modules, which allow critical information to be prioritized and compressed effectively. This demonstrates GSCOOP's practical potential under real-world bandwidth constraints.

Comparisons of Inference times. GSCOOP achieves efficient inference with low latency, scaling from 48.92 ms with one vehicle to 242.35 ms with seven vehicles. Compared to CoBEVT, which increases from 89.11 ms to 511.32 ms over the same range, GSCOOP reduces inference time by nearly 50% and scales efficiently with the number of vehicles, enabling real-time collaborative perception without complex fusion overhead.

4.4. Ablation Study

Component analysis. Table.3 investigates the contribution of several key designs: Relevance-Projection Analysis(RPA), Semantic Uncertainty Evaluation(SUE) and Semantic Gaussian Compression(SGC). The experiment was conducted under the condition of selecting key gaussians, when both RPA and SUE are combined, segmentation accuracy improves significantly across all category. The inclusion of SGC, while resulting in a minimal decrease in segmentation performance, leads to a 53.5% reduction in communication volumes. This highlights the efficiency of our selection and compression modules.

Time delay analysis: We further examine the impact of time delay ranging from 0 to 500ms on OPV2V dataset as shown in Fig.5. Our method exceeds No Fusion's average

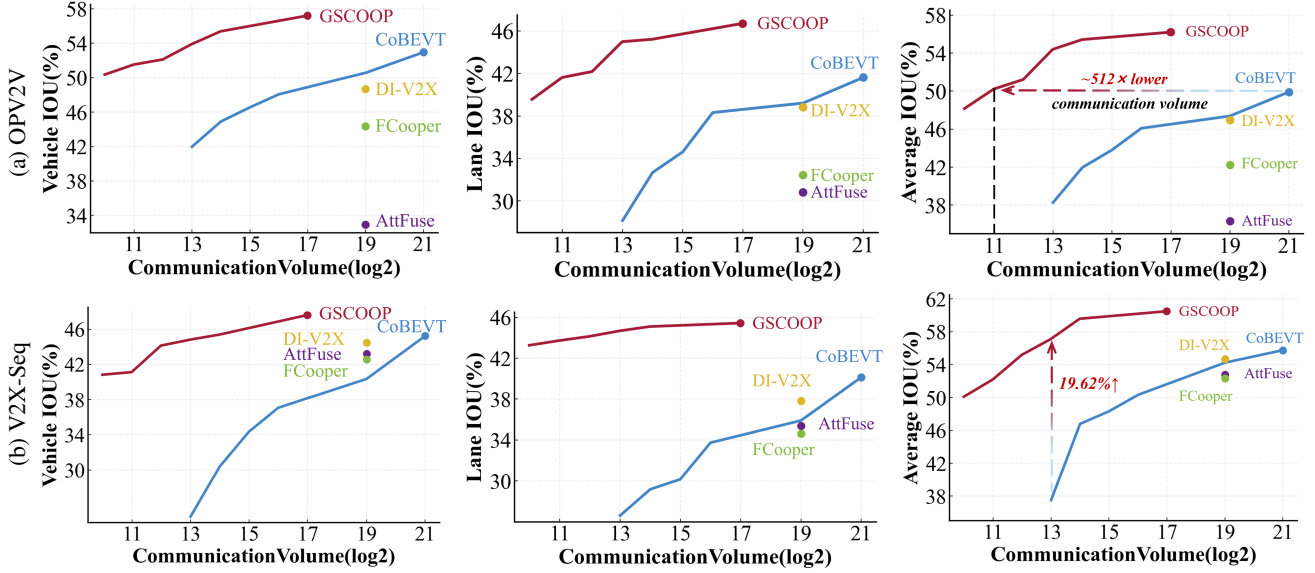


Figure 4. Collaborative semantic segmentation performance comparison of GSCOOP and CoBEVT[38] on the (a) OPV2V and (b) V2X-Seq datasets with varying communication volume.

Table 3. Component Ablation study on the OPV2V dataset.

RPA	SUE	SGC	Veh./Road/Lane \uparrow	Vol.(KB) \downarrow
\checkmark			56.84/64.56/46.32	34.18
	\checkmark		55.20/63.57/45.54	34.18
\checkmark	\checkmark		57.39/64.99/47.10	34.18
\checkmark	\checkmark	\checkmark	57.21/64.77/46.71	15.87

IOU by 16.65% even at 500ms delay. This may be because 3D Gaussian inherently models positional uncertainty, representing probability distributions over finite volumes instead of single points. This spatial spread naturally mitigates time delays by absorbing perturbations, providing GSCOOP with inherent robustness.

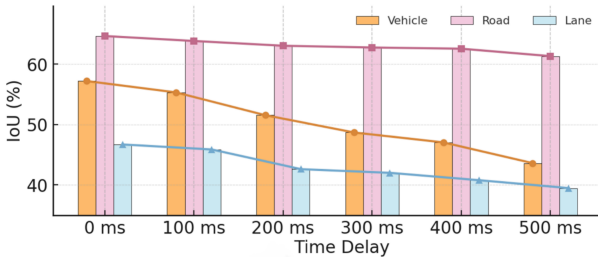


Figure 5. Time delay robustness analysis experiment

Effect of initialization voxel size. We study the impact of voxel size initialization by varying grid dimensions from $80 \times 80 \times 4$ to $120 \times 120 \times 8$. Smaller voxel sizes improve resolution but increase computational costs, while larger sizes reduce precision. GSCOOP performs best with a voxel size of $100 \times 100 \times 4$, as shown in Table.4, which demonstrates the importance of selecting the right voxel size to optimize both performance and resource usage.

Table 4. Ablation study on the effect of initialization voxel size and code book size on all the datasets.

Designs	OPV2V IOU(%)		V2X-Seq IOU(%)	
	Vehicle	Avg	Vehicle	Avg
Effect of Initialization Voxel Size §3.2				
80*80*4	53.62	52.37	43.63	54.63
80*80*8	52.30	51.86	41.41	52.88
100*100*4 (Default)	57.21	56.20	47.61	60.49
100*100*8	55.33	53.82	44.34	57.22
120*120*4	56.03	54.15	44.78	57.44
Effect of Codebook Size §3.4				
8	53.31	52.51	42.48	56.41
32	55.32	53.58	44.31	57.33
64 (Default)	57.21	56.20	47.61	60.49
128	56.68	56.95	45.12	58.15
256	54.79	54.36	43.56	56.54

Effect of codebook size. To evaluate the impact of codebook size on segmentation performance, we experimented with different sizes ranging from 8 to 256. Larger codebooks improve performance but increase overhead and hard to optimize, while smaller ones are more efficient but may miss details. Our experiments suggest that a codebook size of 64 achieves the best trade-off as shown in Table.4, ensuring efficient and accurate representation.

4.5. Qualitative Evaluation

Fig. 6 shows a qualitative comparison across various challenging scenarios. GSCOOP significantly outperforms other SOTA methods[16, 38] in both scene accuracy and detail.

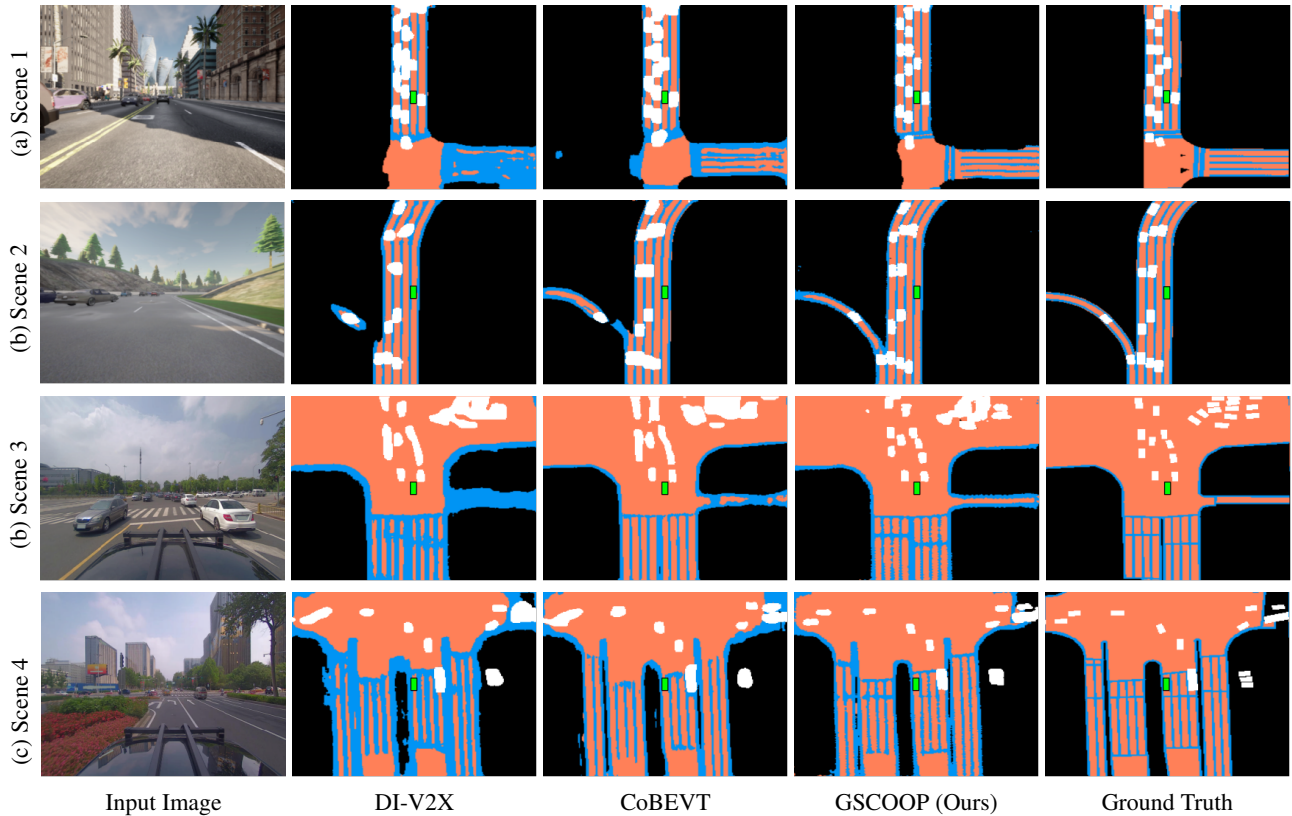


Figure 6. Qualitative comparison segmentation performance across four different scenes. The front view input image, predictions from DI-V2X, CoBEVT, and GSCOOP (ours), along with ground truth annotations, are displayed for each scene. GSCOOP is shown to perform well even under occlusion-heavy conditions and accurately captures detailed segmentation information.

Especially in complex driving scenarios, GSCOOP demonstrates its ability to capture fine-grained scene details through the powerful modeling capability of Gaussian representations and the completeness of spatial information. However, we also observe some limitation in highly dense areas, such as certain regions in Scene 3, there are some ambiguous predictions remain despite outperforming other methods. This may be due to a lack of differentiation during ray sampling of features. Future work may need to explore Gaussian generation schemes from images to address these fine-grained challenges.

5. Conclusion and Limitations

In this paper, we propose GSCOOP, a novel communication-efficient collaborative semantic segmentation framework. Our core idea is to achieve communication-efficient collaborative semantic segmentation by sharing sparse yet accurate 3D Gaussian representation. By selectively extracting, filtering, and compactly encoding Gaussians, GSCOOP effectively reduces bandwidth consumption while maintaining high BEV semantic segmentation performance. Comprehensive experiments on OPV2V and V2X-Seq demonstrate

that GSCOOP achieves state-of-the-art performance, striking an impressive tradeoff between segmentation accuracy and communication efficiency, making it well-suited for bandwidth-constrained collaborative perception systems.

Limitations. While GSCOOP demonstrates significant improvements in collaborative perception, there are still challenges to address in real-world scenarios. Localization errors, communication lossy, and malicious attacks can degrade the system’s performance, particularly in dynamic or unreliable environments. Additionally, the current Gaussian representation could be further optimized to improve both efficiency and accuracy through methodological innovations or by incorporating multi-modal data, such as point clouds that have rich prior position knowledge. In future work, we will focus on addressing these limitations to further boost the system’s performance and resilience.

Acknowledgement. This work is support in part by the National Science Fund for Distinguished Young Scholars (No.62325208), in part by the National Natural Science Foundation of China (No.62232002), and in part by the Natural Science Foundation of Tianjin (No.23ZGZNGX00020 and 24JCZJJC00050).

References

- [1] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3): 1852–1864, 2020. 2
- [2] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. *arXiv preprint arXiv:2407.14108*, 2024. 2
- [3] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, pages 88–100, 2019. 6
- [4] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 514–524. IEEE, 2019. 1
- [5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 5
- [6] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024. 2
- [7] Xiangbo Gao, Runsheng Xu, Jiachen Li, Ziran Wang, Zhiwen Fan, and Zhengzhong Tu. STAMP: Scalable task- and model-agnostic collaborative perception. In *The Thirteenth International Conference on Learning Representations*, 2025. 1
- [8] Haodi He, Colton Stearns, Adam W Harley, and Leonidas J Guibas. View-consistent hierarchical 3d segmentation using ultrametric feature fields. In *European Conference on Computer Vision*, pages 268–286. Springer, 2025. 2
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [10] Chunyong Hu, Hang Zheng, Kun Li, Jianyun Xu, Weibo Mao, Maochun Luo, Lingxuan Wang, Mingxia Chen, Kaixuan Liu, Yiru Zhao, et al. Fusionformer: A multi-sensory fusion in bird’s-eye-view and temporal consistent transformer for 3d objection. *arXiv preprint arXiv:2309.05257*, 2023. 1
- [11] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022. 2
- [12] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. Collaboration helps camera overtake lidar in 3d detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2023. 2
- [13] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15481–15490, 2024. 1
- [14] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pages 376–393. Springer, 2025. 2
- [15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 5
- [16] Xiang Li, Junbo Yin, Wei Li, Chengzhong Xu, Ruigang Yang, and Jianbing Shen. Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3208–3215, 2024. 2, 6, 7
- [17] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 1
- [18] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [20] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, and Xinge Zhu. Vision-centric bev perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [21] Ruiqing Mao, Haotian Wu, Yukuan Jia, Zhaojun Nan, Yuxuan Sun, Sheng Zhou, Deniz Gündüz, and Zhisheng Niu. Diffcp: Ultra-low bit collaborative perception via diffusion model. *arXiv preprint arXiv:2409.19592*, 2024. 1
- [22] Khan Muhammad, Tanveer Hussain, Hayat Ullah, Javier Del Ser, Mahdi Rezaei, Neeraj Kumar, Mohammad Hiji, Paolo Bellavista, and Victor Hugo C de Albuquerque. Vision-based semantic segmentation in scene understanding for autonomous driving: Recent achievements, challenges, and outlooks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22694–22715, 2022. 1
- [23] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird’s eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 1
- [24] Simon Niedermayr, Josef Stumpfegger, and Rüdiger Westermann. Compressed 3d gaussian splatting for accelerated novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10349–10358, 2024. 5
- [25] Lang Peng, Zhirong Chen, Zhangjie Fu, Pengpeng Liang, and Erkang Cheng. Bevsegformer: Bird’s eye view semantic segmentation from arbitrary camera rigs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5935–5943, 2023. 1
- [26] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and

- Ziyan Wu. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*, 2024. 2
- [27] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3
- [28] Donghao Qiao and Farhana Zulkernine. Cobefusion: Cooperative perception with lidar-camera bird’s-eye view fusion. *arXiv preprint arXiv:2310.06008*, 2023. 1, 2
- [29] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 2, 5
- [30] Jiayao Tan, Fan Lyu, Linyan Li, Fuyuan Hu, Tingliang Feng, Fenglei Xu, Zhang Zhang, Rui Yao, and Liang Wang. Dynamic v2x perception from road-to-vehicle vision. *IEEE Transactions on Intelligent Vehicles*, pages 1–14, 2024. 1
- [31] Binglu Wang, Lei Zhang, Zhaozhong Wang, Yongqiang Zhao, and Tianfei Zhou. Core: Cooperative reconstruction for multi-agent perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8710–8720, 2023. 1, 2
- [32] Tianhang Wang, Fan Lu, Zehan Zheng, Guang Chen, and Changjun Jiang. Rcdn: Towards robust camera-insensitivity collaborative perception via dynamic feature-based 3d neural modeling, 2024. 1
- [33] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 605–621. Springer, 2020. 5
- [34] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *arXiv preprint arXiv:2406.02058*, 2024. 2, 5
- [35] Hao Xiang, Runsheng Xu, and Jiaqi Ma. Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 284–295, 2023. 1
- [36] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 1
- [37] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 2, 5, 6
- [38] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. Cobevt: Cooperative bird’s eye view semantic segmentation with sparse transformers. In *Conference on Robot Learning*, pages 989–1000. PMLR, 2023. 1, 2, 5, 6, 7
- [39] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023. 2
- [40] Shaoqing Xu, Fang Li, Shengyin Jiang, Ziyang Song, Li Liu, and Zhi-xin Yang. Gaussianpretrain: A simple unified 3d gaussian representation for visual pre-training in autonomous driving. *arXiv preprint arXiv:2411.12452*, 2024. 2
- [41] Yunjiang Xu, Lingzhi Li, Jin Wang, Benyuan Yang, Zhiwen Wu, Xinhong Chen, and Jianping Wang. Codytrust: Robust asynchronous collaborative perception via dynamic feature trust modulus. *arXiv preprint arXiv:2502.08169*, 2025. 1
- [42] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 1
- [43] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023. 2, 5
- [44] Zhenxun Yuan, Xiao Song, Lei Bai, Zhe Wang, and Wanli Ouyang. Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 2068–2078, 2021. 1
- [45] Jingyu Zhang, Yilei Wang, Lang Qian, Peng Sun, Zengwen Li, Sudong Jiang, Maolin Liu, and Liang Song. Dsrc: Learning density-insensitive and semantic-aware collaborative representation against corruptions. *arXiv preprint arXiv:2412.10739*, 2024. 1
- [46] Jingyu Zhang, Kun Yang, Yilei Wang, Hanqi Wang, Peng Sun, and Liang Song. Ermvp: Communication-efficient and collaboration-robust multi-vehicle perception in challenging environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12575–12584, 2024. 1, 6
- [47] Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024. 2