CoVoT: Communication-Efficient Collaborative Perception via Sparse Voting Transformer

Anonymous Author(s)
Submission Id: 7186

Abstract

Cooperative perception in current vehicular systems faces critical challenges in the trade-off between communication efficiency and perception performance, particularly when handling sparse feature representations across distributed vehicles. To address this issue, we have proposed **CoVoT**, a novel **Co**llaborative **Vo**ting **T**ransformer that fundamentally revisits sparse collaboration through dynamic attention window voting, enabling efficient cross-vehicle collaboration while fully exploiting the inherent sparsity in the point cloud modality. To the best of our knowledge, CoVoT is the first collaboration framework based on fully sparse feature processing in the field of collaborative perception. Specifically, we design a fully sparse transformer integrated with grid-aligned sparse convolution networks to extract sparse voting features (SVFs). During collaboration, vehicles exclusively exchange SVFs to evaluate cross-agent consensus, through which they deterministically identify critical spartial collaboration windows with ultra-low communication overhead while preserving abundant spatial information. The ego vehicle then decodes the received SVFs into window-aligned spatial priors and conducts an attention-based fusion, achieving comprehensive environmental perception through sparsity-consistent feature integration. Experiments on OPV2V, V2XSet and DAIR-V2X demonstrate that **CoVoT** consistently outperforms previous methods in terms of the communication-perception trade-off and spatial robustness. Source code and pre-trained models will be released to the community.

CCS Concepts

• **Computing methodologies** → **Object detection**; **Cooperation and coordination**; *Multi-agent systems*; • **Information systems** → *Sensor networks*.

Keywords

Collaborative perception, Object detection, Intermediate fusion, Sparse, Multi-agent learning

ACM Reference Format:

Anonymous Author(s). 2018. CoVoT: Communication-Efficient Collaborative Perception via Sparse Voting Transformer. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'ACM MM)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'ACM MM, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Autonomous driving has witnessed rapid progress in recent years, driven by advances in sensor technology and deep learning-based algorithms. Despite significant advances in perception algorithms, autonomous vehicles operating in a single-agent setting still face fundamental limitations due to restricted sensor range, occlusions, and incomplete scene understanding. To address these challenges, Collaborative Perception enables individual agents to exchange their perceptual information and fully fuse multi-view observations from surrounding traffic participants, thereby constructing a more comprehensive, wide-area, and robust environmental representation [2, 10, 18, 22–24, 35]. Such cooperation mitigates blind spots and partial occlusions inherent in single-vehicle perception, providing a stronger foundation for autonomous driving systems to make decisions and plan trajectories in complex scenarios.

Nevertheless, in practical applications, balancing perception performance improvements with communication overhead remains a critical challenge for collaborative perception. Existing methods primarily fall into two categories: intermediate fusion approaches [12, 14] which reduce data transmission bandwidth via spatial confidence maps or shared codebooks; and hybrid fusion methods [1], which compress data by replacing partial raw point clouds with high-confidence detection results.

Although these methods alleviate communication burdens to some extent, they still encounter several problems, as illustrated in Figure 2. First, regarding Feature Representation, conventional communication and fusion approaches predominantly rely on transmitting dense intermediate feature maps. While some methods implement selective transmission strategies, their inherently dense representations still contain substantial redundancies that challenge the strict 27MB/s bandwidth limitation specified by V2X communication standards [8, 15]. Second, during Feature Processing and Encoding, existing sparse processing techniques remain constrained by conventional dense detection frameworks, as their features retain dense characteristics at the core. Current sparse transmission methods [12, 14, 28, 31, 33] employ brute-force feature truncation strategies that compromise perception accuracy for communication efficiency, failing to fully exploit the inherent advantages of sparse representations in computational and communication cost reduction. Third, in the Communication and Fusion stage, conventional approaches typically apply post-hoc sparsification through confidence thresholds or request maps after processing full dense features, often resulting in critical spatial information loss during

mask truncation. This limitation not only degrades perception accuracy but also reduces system robustness against the environmental noise.

To overcome the above challenges, we propose a novel collaborative perception framework — **CoVoT** (Collaborative Voting Transformer). Unlike conventional methods that first process dense features and then perform post-hoc sparsification, CoVoT is designed to operate on **sparse feature representations**, enabling fine-grained processing and optimization directly on the sparse data. On this basis, during subsequent communication processes, we can preserve as many scene features as possible instead of truncating or discarding them. In this context, we employ the concept of voting, the core of which is to establish consensus among individual features. During the later stages of communication, we further fuse the features based on the single-view consensus obtained through voting.

Specifically, this paper first introduces an efficient sparse feature representation method that directly extracts the sparse features from point clouds using deep sparse networks, effectively avoiding the redundancy introduced by dense feature processing and fully leveraging the advantages of sparse representations in low-bandwidth communication scenarios. Secondly, to resolve information truncation issues in existing feature selection and encoding methods, we propose a voting encoder/decoder based on sparse convolution. This architecture implements precise spatial information encoding through region-based voting mechanisms and neighborhood-aware convolution, ensuring that transmitted messages contain compact yet spatially robust features. Moreover, in the communication and fusion stage, the proposed Voting Information Confirmation (VIC) mechanism utilizes cross-vehicle voting strategies to further filter and fuse the compressed features generated during encoding based on region-specific confidence weights. This approach effectively recovers and supplements key spatial details, thereby achieving higher perception accuracy under stringent communication bandwidth constraints.

The main contributions of this paper can be summarized as follows:

- We propose the **CoVoT** framework, a novel and efficient multi-agent collaborative perception framework that achieves state-of-the-art (SOTA) detection accuracy at a minimal communication cost. To the best of our knowledge, CoVoT is the first collaboration framework based on fully sparse feature processing in the field of collaborative perception.
- We revisit critical challenges in collaborative perception communication pipelines, addressing key issues in feature representation, transmission, and fusion. By designing a fully sparse feature processing pathway and transmission mechanism, we establish a paradigm that guides multi-agent to exchange the sparse features from regions that reach a voting consensus.
- Experimental results demonstrate that, under extremely low communication overhead, CoVoT significantly outperforms mainstream methods by achieving SOTA-level collaborative detection performance while exhibiting strong spatial robustness.

2 Related Works

2.1 Data Fusion in Collaborative Perception

Collaborative perception improves perception performance by sharing information among multiple agents. With the development of single-vehicle perception algorithms, visual language models (VLMs), and other technologies, it is playing an increasingly important role in the field of autonomous driving. Collaborative perception can be divided into three categories according to the stage of integration: Early fusion, intermediate fusion, and late fusion. **Early fusion** retains the original modality and interpretability, preserving full visual and 3D information. [25] builds a fully connected communication network and [16] utilizes knowledge distillation on early fusion information to design an object-specific intermediate fusion algorithm. V2X-VLM [32] integrates text prompts and camera images for collaboration, pioneering the combination of collaborative perception and VLMs for autonomous driving planning tasks. **Late fusion**, on the other hand, only transmits perception results for collaboration, achieving efficient communication but at the cost of losing critical information, leading to weaker collaboration performance and robustness.

The core challenge of collaborative perception lies in designing efficient collaboration mechanisms to balance perception accuracy and communication overhead. Based on collaborative perception datasets such as [27, 29, 34], mainstream algorithms tend to adopt **intermediate fusion** approaches to achieve a better performance-communication trade-off [12, 14, 31, 33]. When2comm [19] introduces a handshake mechanism to select communication agents, V2X-ViT [28] designs a unified transformer to capture inter-agent interaction and spatial relationship, Where2comm [12] transmits features to specific locations based on spatial confidence, How2comm [31] further incorporates a temporal prediction model, Pragcomm [13] alternates between transmitting local dynamic and global features, CodeFilling [14] maintains a codebook among agents and transmits integer codes, Which2comm [33] obtains sparse objects features for transmission using a sparse network and achieves feature fusion with a relative temporal encoding mechanism.

However, most of these methods are based on dense BEV feature maps for processing and communication. Although they achieve good performance, they still have significant communication and computation overheads, especially in multi-agent and complex environments. To solve this problem, we introduce sparse convolution to perform computation and information transfer only in effective regions. At the same time, the sparse feature representation is naturally suitable for sparse inputs such as point clouds, which can retain finer structural information. Combined with the sparse convolution cooperative perception framework, it not only maintains or even improves perception accuracy under communication constraints but also improves the overall computational efficiency and scalability of the system.

2.2 Sparse Convolution

Sparse convolution is a convolution operation specifically designed to handle sparse data and is commonly used in point cloud processing [22, 23], autonomous driving perception[36] and 3D object

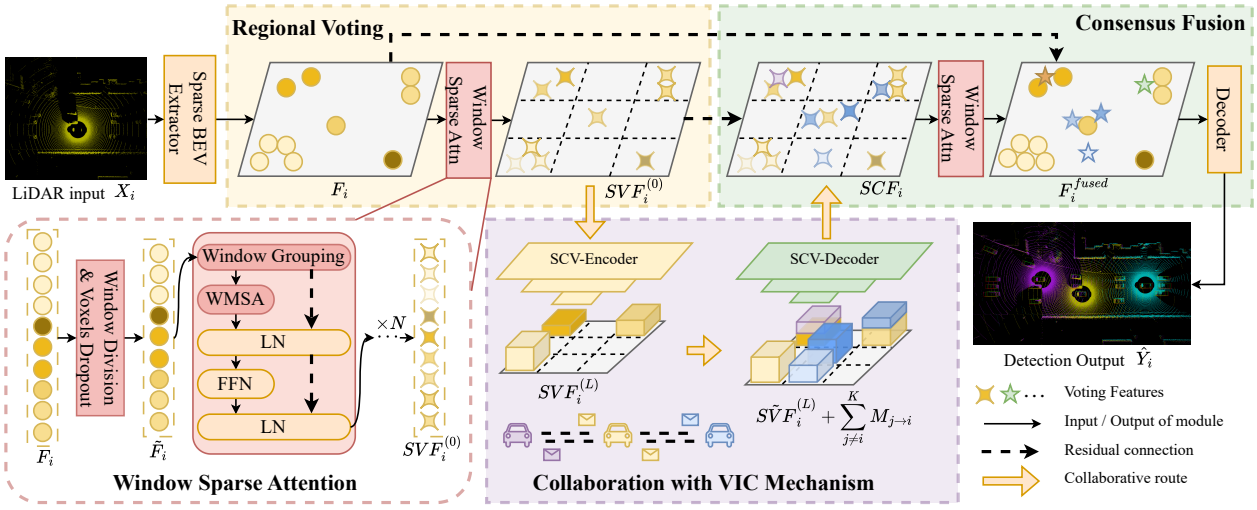


Figure 1: Framework of CoVoT, a Collaborative Voting Transformer with ultra-low communication cost, consisting of a *sparse feature extraction* and a *convolution-integrated sparse transformer (CIST)*. Ego SVFs are generated from the sparse BEV features through *WSA* and down-sampled by *SCV Encoder*, received SVFs are decoded and concatenated with ego SVFs before the consensus fusion, where features can be globally enhanced, and further processed through Decoder.

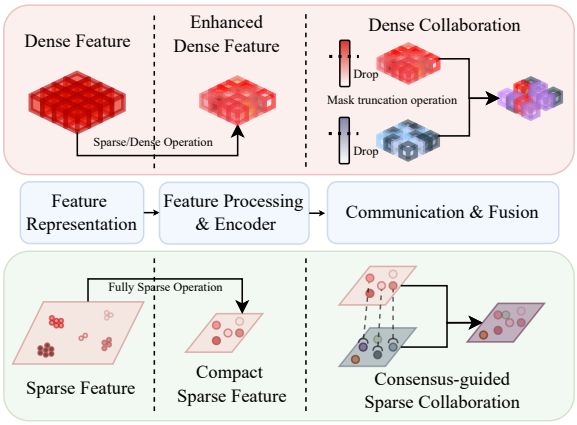


Figure 2: Comparison between existing collaborative perception frameworks (top) and our proposed framework (bottom). The feature communication pipeline comprises three key stages: 1) Feature representation, 2) Feature processing & encoding, and 3) Feature communication & fusion. Existing approaches [12, 14, 28, 31, 33] sequentially process dense features followed by post hoc sparsification via confidence or request maps, inevitably discarding critical spatial information during mask truncation. In contrast, our framework maintains and refines fully sparse representations throughout all processing stages.

detection [30, 37]. Traditional dense convolution performs computations on the entire input feature map or voxel grid, even if most of the regions are empty, leading to computational inefficiency. Sparse convolution, on the other hand, greatly reduces computational overhead and storage requirements by performing computations only at valid locations, enabling models to process large-scale 3D data more efficiently. VoxelNeXT [3] combines efficient voxel feature extraction and sparse convolution to reduce computational cost while maintaining high performance. SST [6] employs a sparse attention mechanism to perform local and global feature interactions on point cloud data, which further improves detection accuracy and computational efficiency. FSD [7] explores a fully sparse 3D target detection method that performs computation and inference only in sparse locations and shifts edge points to center of objects, thus significantly reducing computational requirements while maintaining competitive detection performance.

Inspired by [3] and [6], we compute sparse features to dynamically vote for the selection of sparse feature groups, enabling effective collaboration across sparse features. Preserving the inherent sparsity of the LiDAR modality, we introduce a voting mechanism based on learned features that estimate the importance of inter-window interactions. These voting features act as attention signals that guide the model in identifying which windows should cooperate, thereby enhancing spatial awareness while maintaining computational efficiency.

3 Problem Formulation

Consider a system with K connected autonomous vehicles (CAVs). Let X_i denote the point cloud data captured by vehicle i, and Y represent the ground truth for cooperative perception. In the collaborative 3D detection pipeline as shown in Fig.1, The complete

workflow for vehicle i can be formalized as:

$$F_i = \Phi_{\text{enc}}(X_i) \tag{1}$$

$$M_{i\to j} = \Phi_{\text{comp}}(F_i, M_{j\to i}) \tag{2}$$

$$\mathcal{E}_i = g\left(\Phi_{\text{dec}}(\Phi_{\text{fuse}}(F_i, \{M_{j\to i}\}_{j=1,j\neq i}^{K})), Y\right) \tag{3}$$

where F_i denotes encoded features, $M_{j\to i}$ represents compressed collaborative messages, and $g(\cdot,\cdot)$ measures the performance of perception of the ith CAV to be $\mathcal{E}_i$; $\Phi_{\text{enc}}(\cdot)$,$\Phi_{\text{comp}}(\cdot)$, $\Phi_{\text{fuse}}(\cdot)$ and $\Phi_{\text{dec}}(\cdot)$, respectively, denote operators representing the extraction process, the message compressing process, the feature fusion process and detection process. To collaboratively improve performance, vehicles exchange collaborative information through the following dual-role mechanism:

Role of the Ego Vehicle: When acting as the ego vehicle, each agent i aggregates relevant information from collaborators and fuses it with local observations:

$$\hat{Y}_i = \Phi_{\text{dec}}\left(\Phi_{\text{fuse}}\left(F_i, \{M_{j\to i}\}_{j=1,j\neq i}^{K}\right)\right) \tag{4}$$

Roles of Collaborative Vehicles: When serving other vehicles, each agent i generates compact collaborative messages through importance-aware compression:

$$M_i = \Phi_{\text{comp}}\left(\Phi_{\text{enc}}(X_i), M_{j\to i}\right) \tag{5}$$

Due to the limitation of communication volume in the scenario of collaborative perception, the system-wide objective under bandwidth constraint B is formulated as follow:

$$\max_{\Phi_{\text{enc}},\Phi_{\text{comp}},\Phi_{\text{fuse}},\Phi_{\text{dec}}} \sum_{i=1}^{K} \mathcal{E}_i \quad \text{s.t.} \quad \sum_{i=1}^{K} \mathcal{V}(M_i) \leq B \tag{6}$$

where $\mathcal{V}(\cdot)$ calculates the communication volume of messages.

4 CoVoT: Intermediate Collaboration Framework Based on Sparse Transformer

In this section, we propose **CoVoT**, a sparse feature voting fusion framework that optimizes the perception-communication trade-off through sparse attention and collaborative voting. The novelty of architecture consists of two modules: *sparse feature encoding* and *convolution-integrated sparse transformer (CIST)*, which will be introduced in Section 4.2 and Section 4.3. In addition, a new collaboration mechanism *Voting Information Confirmation (VIC)*, is proposed and introduced in Section 4.4.

4.1 Overall architecture

As illustrated in Figure 1, the CoVoT framework processes point cloud data through three key stages: Firstly, a *sparse feature encoder* extracts the sparse features F_i of raw point clouds X_i to implement Equation (1). Subsequently, these features are then further processed by the *CIST* module via: 1) *regional voting with* the *window sparse attention (WSA)* [20] that balances the features via window division, feature dropout and self-attention, 2) *sparse convolution voting (SCV) encoder* that aggregates votes from neighboring window to generate the window-level *sparse voting features (SVFs)* with a highly coarse spatial resolution, 3) *SCV decoder* with transposed sparse convolution that decodes the collaborative *SVFs* into high-resolution sparse features and 4) *consensus fusion* where ego vehicle utilizes *WSA* again motivate the regional fusion between ego *SVFs* and collaborative *SVFs* that reached consensus.Finally, the detection head decodes the integrated features to produce detection results $\hat{Y}_i$. In this architecture, the collaboration messages are exchanged after where collaborative agents exchange these highly coarse *SVFs* via the *VIC* mechanism, through which the ego vehicle will be able to selectively transmit spatially critical voting features that reach consensus with collaborative CAVs.

4.2 Sparse Feature Extraction

In this module, a voxel-based feature extraction method [3] is adopted to process the point cloud into sparse features. The point cloud X_i is encoded into structured voxels that record the mean features of the points in each voxel, denoted $V_i \in \mathbb{R}^{n\times c}$ with their corresponding indices $I_{v,i} \in \mathbb{R}^{n\times 4}$, where n denotes the number of voxels and c denotes the number of feature channels.

After the basic mean voxel feature encoding, the sparse extractor, comprising multi-scale sparse convolution blocks, operates on the encoded voxel features to dynamically capture the relationship between feature points while preserving the sparsity of the point-cloud modality. In each layer of the block, focal loss [17] supervision is adopted to supervise the relationship of locations between features extracted and the ground truth bounding-boxes, ensuring that feature points sparsely distributed inside the boxes. The results of the ablation analysis demonstrate that the sparse feature extractor is an effective strategy to enhance performance. The overall feature extraction module serves as the feature encoder mentioned in Equation (1), where the sparse BEV features $F_i \in \mathbb{R}^{n_f\times c_f}$ and their corresponding indices and $I_i \in \mathbb{R}^{n_f\times 4}$ are extracted from the original Voxel features $(V_i, I_{v,i})$ as follow:

$$(F_i, I_i) = \Phi_{\text{ext}}(V_i, I_{v,i})) \tag{7}$$

where n_f and c_f denote the number of encoded feature points and channels, respectively.

4.3 Convolution-Integrated Sparse Transformer

This module integrates sparse convolution blocks into the sparse transformer architecture to achieve two objectives: 1) enhanced feature refinement for the sparse features extracted in Section 4.2, and 2) cross-agent feature fusion. For sparse feature processing, global self-attention often becomes computationally prohibitive due to memory constraints, while sparse detection targets primarily correlate with their local features. Although windowed attention strategies are typically adopted for efficient computation, cooperative perception in vehicular networks faces additional challenges: incomplete sparse features within local windows and the substantial communication overhead required for transmitting collaborative spatial information. To address these issues, we propose a Convolution-Integrated Sparse Transformer that encodes, utilizes and decodes SVFs through three synergistic components: Regional Voting with Self-Attention, Sparse Convolution Voting Encoder, and Cross-Agent Voting Enhancement.

4.3.1 Regional Voting & Sparse Convolution Voting Encoder. Given the sparse BEV features (F_i, I_i) distributed across the voxel space of dimension $H\times W$, we partition them into $H_w \times W_w$ windows for group-wise processing. The *WSA* [20] is applied to enhance local

voting features within each partition. To mitigate boundary effects caused by fixed window partitioning, where adjacent features might be separated into different windows, we implement shifted window attention with $(H_w/2, W_w/2)$ offsets. This staggered windowing strategy reinforces local feature interactions through overlapping receptive fields.

Specifically, we pre-process (F_i, I_i) into $(\tilde{F}_i, \tilde{I}_i)$ employ adaptive *indices dropout* to dynamically balance the number of features that participate in self-attention within each window. This mechanism reduces computational complexity and helps balance the feature distribution of the whole BEV space, contributing to enhancing the saliency of voting features from regions that need collaboration from other CAVs. The following voting feature generation and enhancement can be formulated as:

$$\tilde{F'}_i = WMSA(LN(F_i)), PE(\tilde{I}_i) + \tilde{F}_i \tag{8}$$

$$\tilde{F}_i = FFN(LN(\tilde{F'}_i)) + \tilde{F'}_i \tag{9}$$

where $WMSA$ denotes windowed multi-head self-attention [20], LN layer normalization, PE positional encoding preserving global spatial awareness, and FFN the feed-forward network. The resulting voting features leverage sparsity patterns to encapsulate localized contextual information, making them more representative for cooperative feature selection.

The *sparse convolution voting (SCV)* encoder blocks in the *CIST* module hierarchical aggregates of cooperative voting features into coarser representations. This process helps lower the communication cost by down-sampling the SVFs and meanwhile provides robustness against spatial misalignment caused by localization offsets among collaborative vehicles without the compensation on extrinsic parameters. By implementing dilated sparse convolutions with adaptive receptive fields, this component progressively gathers detailed SVFs into a more compact voxel space that properly indicates the voting results of each attention group in *WSA* [20] with a lower feature dimension, maintaining both communication and computational efficiency through sparse tensor operations. After each gathering convolution block that down-samples the SVFs, submanifold sparse convolution layers [9] are applied to expand the gathering range and to maintain the indices of each voting point, while residual connection [11] structures come to eliminate degradation. Finally, the encoded compact SVFs are obtained through Equation (10):

$$SVF_i^{(l+1)} = \mathcal{S}pconv_d(SVF_i^{(l)}) + \mathcal{S}pconv_s(\mathcal{S}pconv_d(SVF_i^{(l)}) \tag{10}$$

where $\mathcal{S}\text{pconv}_{d/s}(\cdot)$ denotes the down-sampling sparse convolution block and the submanifold sparse convolution block, respectively. l denotes the lth layer of the sparse convolution from 1 to L and $SVF_i^{(0)}$ specifically represents the output of the self-attention module. After the convolution layer encoding process, $SVF_i^{(L)}$ will be shared among CAVs. The concrete mechanism of communication will be introduced in Section 4.4.

4.3.2 Sparse Convolution Voting Decoder & Consensus Fusion. The *SCV* decoder receives $S\tilde{V}F_j^{(L)}$, a weighted selection of $SVF_j^{(L)}$ that represents the consensus regions. The decoder layers will reconstruct high-resolution collaborative features for the ego CAV i through the concatenation of $SVF_i^{(L)}$ and $S\tilde{V}F_j^{(L)}$ of all collaborative CAVs. Note that the reconstructing up-sampling architecture is symmetric to the encoder, outputting the *sparse consensus features (SCF)*. This module enables collaborative features highly adaptable to position noise as a result of the ambiguous down-sampling process in the encoder.

To enable more effective and robust collaboration, consensus fusion utilizes the *WSA* module [20], but with the input consensus features from collaborative CAVs, the attention here will aggregate decoder consensus information and thus emphasize the critical features needed in the perception task. After *WSA*, a residual connection [11] from the extracted BEV features (F_i, I_i) is made. Throughout *CIST*, we are trying to enhance features from detection-poor regions but suppress features of high saliency. As a result, this residual connection is also critical to finally boost the performance, sending the F_i^{fused} to the detection decoder.

4.4 Voting Information Confirmation

An *MLP* is introduced in the *VIC* stage to calculate a weight $W^i = MLP(SVF_i^{(L)})$ that indicates the importance of the voting region during collaboration, which emphasizes collaborative voting regions and meanwhile acts as a handshaking message for the CAVs to confirm voting information. Before CAVs exchange their voting features, W is first transmitted to broadcast their knowledge about the coarse divided regions, and the difference in information is evaluated by a weighted chamfer loss, inspired by [5]:

$$\begin{aligned}\mathcal{L}_{\text{wc}} = &\frac{1}{N}\sum_{n=1}^{N}\min_m\left(\|P_n - P_m\|_2 \cdot \frac{W_n^i + (1 - W_m^j)}{2}\right) + \\ &\frac{1}{M}\sum_{m=1}^{M}\min_n\left(\|P_n - P_m\|_2 \cdot \frac{W_n^i + (1 - W_m^j)}{2}\right)\end{aligned} \tag{11}$$

where P is the normalized position in the coordinates, $n \in [1, \ldots, N]$ and $m \in [1, \ldots, M]$ represent the index of the voted region centers. By introducing confidence-aware weighting into the Chamfer distance computation, our loss achieves dual objectives: i) suppressing propagation of geometrically certain ego-vehicle observations through penalty relaxation, and ii) promoting transmission of complementary spatial features from collaborative partners via gradient-driven attention to high-uncertainty regions.

During collaborative feature fusion, we first sparsify the SVFs by selecting top-K regions guided by W^i, followed by generating compact collaborative messages M_i through weight-adaptive feature aggregation:

$$S\tilde{V}F_i^{(L)} = \sum_{n \in topk(W^i)} W_n^i \cdot (SVF_i^{(L)})_n \tag{12}$$

Thus, the communication volume upper bound of the $SVF_i^{(L)}$ from the ith CAV can be calculated according to the following equation:

$$\mathcal{V}(M_i)_{max} = \underbrace{(N \times 3 \times 4\text{B})}_{\mathcal{V}(W^i)} + \underbrace{(N \times (c^L + 3) \times 4\text{B})}_{\mathcal{V}(S\tilde{V}F_i^{(L)})} \tag{13}$$

where the factor $N_{SVF_i^{(L)}}$ represents the number of voting features from $SVF_i^{(L)}$, which is not greater than $\frac{H}{2^L} \times \frac{W}{2^L}$; the factor $c^L + 3$

Table 1: Performance comparison and communication cost on OPV2V [29], V2XSet [28] and DAIR-V2X [34]. We evaluate our CoVoT under different communication volume, showing SOTA performances among all the tested datasets.

Dataset	OPV2V			V2XSet			DAIR-V2X		
Method/Metric	AP@0.5↑	AP@0.7↑	AB† ↓	AP@0.5↑	AP@0.7↑	AB† ↓	AP@0.5↑	AP@0.7↑	AB† ↓
No Fusion	0.796	0.696	0	0.722	0.588	0	0.512	0.342	0
Early	0.915	0.873	19.59	0.900	0.837	19.81	0.550	0.385	19.09
Late	0.852	0.744	6.59	0.824	0.723	6.04	0.534	0.361	5.44
OPV2V [29]	0.864	0.786	21.62	0.878	0.750	21.62	0.556	0.432	20.92
V2VNet [25]	0.918	0.845	23.04	0.896	0.784	23.04	0.560	0.445	22.33
V2X-Vit [28]	0.907	0.829	23.04	0.901	0.812	23.04	0.573	0.457	22.33
CoAlign [21]	0.921	0.854	23.04	0.908	0.819	23.04	0.580	0.460	22.33
Where2comm [12]	0.884	0.792	21.00	0.876	0.783	20.93	0.552	0.440	19.55
CodeFilling [14] ‡	0.916	0.841	13.92	0.909	0.821	14.56	-	-	-
Ours (CoVoT)	**0.951**	**0.915**	13.93	**0.923**	**0.858**	**13.97**	**0.620**	**0.521**	**13.42**
	0.950	**0.915**	**13.53**	0.921	0.856	13.58	0.614	0.513	12.93
	0.943	0.902	12.83	0.901	0.831	12.88	0.588	0.487	12.31
	0.924	0.876	12.16	0.870	0.790	12.19	0.570	0.471	11.62
	0.906	0.847	11.54	0.843	0.752	11.60	0.565	0.463	10.91
	0.893	0.795	10.15	0.786	0.674	10.30	0.554	0.455	9.74

†: Communication efficiency is also evaluated following the format of [12] denoted as Average Bytes (AB) to calculate the bytes of data transmitted in $\log_2(B)$ scale.
‡: The original implementation is publicly available, but reproduction attempts following the provided instructions failed to achieve the reported performance, with similar issues documented in the repository's community discussions. Nevertheless, we retain their reported results for baseline comparison purposes, with experimental data sourced from [33].

represents the number of output channels in the Lth layer with a 2D spatial indices and a CAV index, while the factor 4B represents the byte-width of the data type. The total communication will be even lower with the *VIC* mechanism to prune the SVFs into sparser ones denoted as $\tilde{SVF}_i^{(L)}$.

5 Experiment Result

5.1 Datasets and Implementation Details

Datasets. To evaluate the effectiveness of CoVoT on the collaborative perception task, experiments are conducted on three widely used multi-agent collaborative perception datasets. **OPV2V** [29] is a large-scale simulated V2V dataset containing over 70 driving scenes and 10,914 annotated LiDAR point cloud frames. **V2XSet** [28] is a simulated V2X dataset generated from CARLA [4] and OpenCDA [26], comprising 11,447 frames in total and includes up to 5 connected agents per scene. **DAIR-V2X** [34] is a real-world dataset containing 100 driving scenarios and 18,000 data samples from vehicles and roadside infrastructure, and we adopt the original dataset that only annotate the ground truth of the intersection for collaborative perception tasks, so that the LiDAR data range will be smaller in the implementation.

Implementation Details. To ensure the objective comparison of various perception methods, the singular LiDAR point cloud modality is utilized as the input for all the methods. The LiDAR range for CAVs in our experiment is $x \in [-140.8m, +140.8m]$, $y \in [-38.4m, +38.4m]$ on OPV2V and V2XSet, and $x \in [0m, +92.16m]$, $y \in [-46.08m, +46.08m]$ on DAIR-V2X. Since we conduct our experiments on the voxel-based backbone, the test methods share a modified SECOND feature extractor where the voxelnet-stage is replaced by the sparse extractor in Section 4.2. The voxel resolution is set to $0.1m \times 0.1m \times 0.1m$ for OPV2V and V2XSet and $0.08m \times 0.08m \times 0.08m$ for DAIR-V2X. Perception performance is evaluated with average precision (AP) at Intersection-over-Union (IoU) 0.5 and 0.7. All models are trained on NVIDIA RTX 3090 GPUs with the Adam optimizer.

5.2 Quantitative Evaluation

Table 1 compares CoVoT with different SOTA intermediate fusion models including OPV2V [29], V2VNet [25], V2X-ViT [28], Where2comm [12], CoAlign [21] and CodeFilling [14]. The results indicate that CoVoT significantly outperforms previous methods in real-world and simulated datasets both in detection performance and communication efficiency, and the experiment on robustness our models outperforms previous SOTAs on the resiliency to localization errors.

5.2.1 Detection Performance Comparison. Looking into the the comparison on the detection performance, our method outperforms previous SOTA V2X-ViT and Where2comm by a significant improvement of 8.56 and 11.35 on OPV2V, respectively, indicating that our voting mechanism not only effectively enhances the features with the sparse transformer than the dense one, but also enables CAVs to reach preciser voting consensus than confidence-based mechanism.

5.2.2 Communication Efficiency Comparison. As shown in Table 1, previous methods such as V2VNet and V2X-ViT achieve high detection performance but require transmission of dense features, leading to significant bandwidth consumption. In contrast, Where2comm partially mitigate this by selecting transmitted features, yet still operate on heavy BEV tensors. CoVoT achieves comparable or superior detection performance while reducing the communication

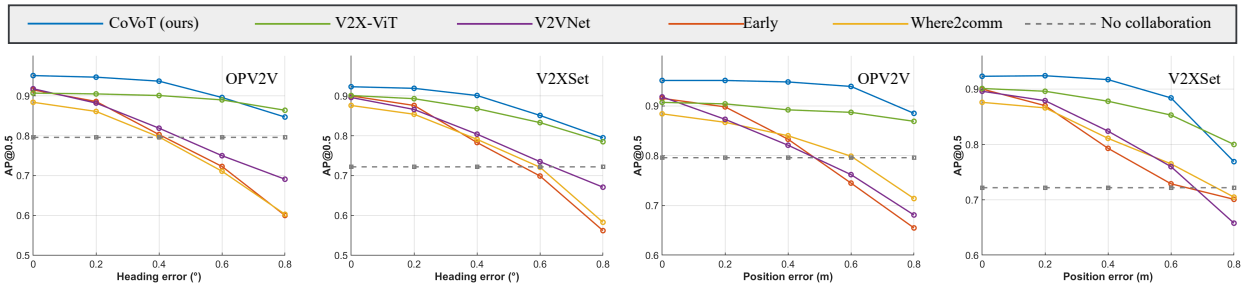
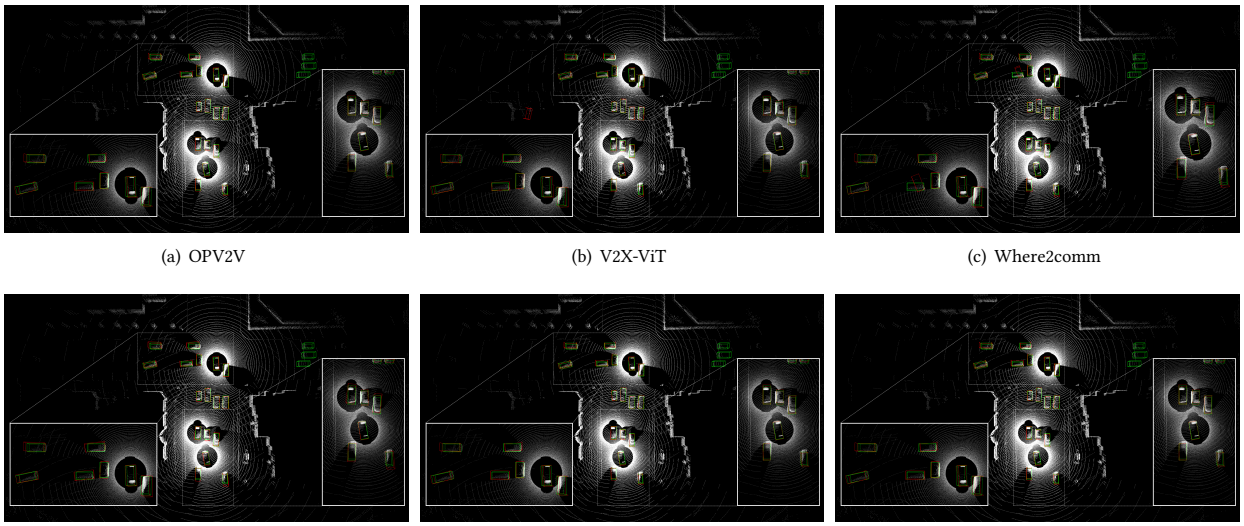


Figure 3: Evaluation of spatial robustness on OPV2V and V2XSet considering position error and heading error.



(a) OPV2V (b) V2X-ViT (c) Where2comm

(d) V2VNet (e) Ours (f) Ours (with noise)

Figure 4: Visualization in intersection scenario from the OPV2V dataset. Green and red boxes represent the ground truth and detection boxes respectively.

volume by more than *14*× comparing with CodeFilling, indicating that the SVFs in CoVoT is well-suited for scalable V2X perception under bandwidth constraints than the code-book based on dense features.

5.2.3 Spatial Robustness. To evaluate robustness, we verify the performance of CoVoT under varying spatial errors shown in Figure 3 following the noise settings in [29]. Experiments are conducted by injecting Gaussian distributed pose perturbations with a standard deviation of $\sigma_{xyz} \in [0, 0.8m]$ and $\sigma_{rpy} \in [0, 0.8°]$ for localization errors of position and heading. The results exhibit the robustness of CoVoT against spatial noises. This is attributed to i) collaboration with *VIC* mechanism transmits voting information that reaches regional consensus; ii) the *SCV Decoder* dynamically reconstruct the voting feature with a rectified location and the consensus fusion ensures that essential spatial information from collaborating agents is seamlessly integrated, while mitigating interference from misaligned or occluded areas; iii) a residual connection from the original sparse BEV ensures the localization of ambiguous regions.

5.3 Qualitative Evaluation

Figure 4 visualizes the detection results of V2X-ViT, Where2comm, V2VNet and CoVoT in intersection scenario from the OPV2V dataset. In scenarios with occlusions, such as intersections, perception performance tends to degrade. Figure 4(a) and Figure 4(b) produce several shifted and oversized boxes, particularly in the lower half of the scene. Figure 4(c)'s detection precision is not sufficiently high due to token-based fusion errors or selection noise. Figure 4(d) also suffers from hallucinations and misalignment. However, as shown in Figure 4(e), compared to the ground truth bounding boxes, our model provides more accurate detection, especially in heavily occluded areas such as the upper half of the intersection far from the ego vehicle. The comparison between Figure 4(e) and Figure 4(f) can also show the robustness of our model.

Table 2: Ablation study results on OPV2V and DAIR-V2X dataset. Abbreviation of modules are used: SE (sparse BEV encoder), RV&CF (regional voting and consensus fusion), VIC (voting information confirmation).

SE	RV&CF	VIC	OPV2V		DAIR-V2X	
			AP@0.5	AP@0.7	AP@0.5	AP@0.7
-	-	-	0.891	0.849	0.532	0.419
✓	-	-	0.923	0.892	0.562	0.466
✓	✓	-	0.944	0.903	0.585	0.483
✓	✓	✓	0.951	0.915	0.620	0.521

5.4 Ablation Studies

To better understand the contribution of each component in CoVoT, we conduct the ablation experiments in Table 2 by removing modules without alternatives structures to replace them. Concretely, the SVFs are directly transmitted and directly concatenated for fusion without the Voting Information Confirmation mechanism, Sparse BEV features are directly input into the decoder without the Region Voting & Consensus Fusion, and dense BEV will be extracted instead of sparse ones without Sparse BEV Extractor. The results indicate that: i) Sparse BEV Extractor can effectively capture more details of the whole sparse scene; ii) Region Voting & Consensus Fusion highlight the role of voting features to boost the collaboration; iii) Voting Information Confirmation mechanism can significantly enhance the performance especially on regions that highly overlapped like the intersection on DAIR-V2X.

6 Conclusion

This paper introduces **CoVoT**, a novel sparse collaboration framework that revisits communication-efficient cooperative perception through the lens of dynamic attention-based window voting. By leveraging grid-aligned sparse convolutions and self-attention mechanisms, CoVoT extracts highly informative sparse voting features (SVFs) that preserve spatial awareness while minimizing redundancy. Through selective SVF exchange and attention-guided fusion, our method achieves accurate and robust multi-agent perception with ultra-low communication cost. Unlike prior approaches that rely on post hoc sparsification of dense features, CoVoT operates directly on sparse representations throughout the entire pipeline, ensuring consistency, geometric precision, and efficiency. Extensive evaluations on the OPV2V, V2XSet and DAIR-V2X datasets validate that our approach significantly outperforms SOTA methods across the communication-accuracy spectrum. Future research will investigate extending our sparse-centric methodology to heterogeneous sensor fusion scenarios, particularly exploring cross-modal synergy between LiDAR, radar, and visual inputs. Such extensions could enable more robust environmental understanding through complementary modality-specific features, ultimately enhancing decision-making robustness for safety-critical autonomous navigation systems.

References

[1] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. 2020. Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems* 23, 3 (2020), 1852–1864.

[2] Antoine Caillot, Safa Ouerghi, Pascal Vasseur, Rémi Boutteau, and Yohan Dupuis. 2022. Survey on cooperative perception in an automotive context. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 14204–14223.

[3] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21674–21683.

[4] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16.

[5] Haoqiang Fan, Hao Su, and Leonidas J Guibas. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 605–613.

[6] Lue Fan, Ziqi Pang, Tianyuan Zhang, Yu-Xiong Wang, Hang Zhao, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. 2022. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8458–8468.

[7] Lue Fan, Feng Wang, Naiyan Wang, and Zhao-Xiang Zhang. 2022. Fully sparse 3d object detection. *Advances in Neural Information Processing Systems* 35 (2022), 351–363.

[8] Alessio Filippi, Kees Moerman, Vincent Martinez, Andrew Turley, Onn Haran, and Ron Toledano. 2017. IEEE802. 11p ahead of LTE-V2V for safety applications. *Autotalks NXP* 1 (2017), 1–19.

[9] Benjamin Graham and Laurens Van der Maaten. 2017. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307* (2017).

[10] Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. 2023. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intelligent Transportation Systems Magazine* 15, 6 (2023), 131–151.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. 2022. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems* 35 (2022), 4874–4886.

[13] Yue Hu, Xianghe Pang, Xiaoqi Qin, Yonina C Eldar, Siheng Chen, Ping Zhang, and Wenjun Zhang. 2024. Pragmatic communication in multi-agent collaborative perception. *arXiv preprint arXiv:2401.12694* (2024).

[14] Yue Hu, Juntong Peng, Sifei Liu, Junhao Ge, Si Liu, and Siheng Chen. 2024. Communication-efficient collaborative perception via information filling with codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15481–15490.

[15] Daniel Jiang, Qi Chen, and Luca Delgrossi. 2008. Optimal data rate selection for vehicle safety communications. In *Proceedings of the fifth ACM international workshop on VehicuAr Inter-NETworking*. 30–38.

[16] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. 2021. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems* 34 (2021), 29541–29552.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

[18] Si Liu, Chen Gao, Yuan Chen, Xingyu Peng, Xianghao Kong, Kun Wang, Runsheng Xu, Wentao Jiang, Hao Xiang, Jiaqi Ma, et al. 2023. Towards vehicle-to-everything autonomous driving: A survey on collaborative perception. *arXiv preprint arXiv:2308.16714* (2023).

[19] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. 2020. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 4106–4115.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[21] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. 2023. Robust collaborative 3d object detection in presence of pose errors. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4812–4818.

[22] Qianxin Qu, Yijin Xiong, Guipeng Zhang, Xin Wu, Xiaohan Gao, Xin Gao, Hanyu Li, Shichun Guo, and Guoying Zhang. 2024. V2I-Calib: A Novel Calibration Approach for Collaborative Vehicle and Infrastructure LiDAR Systems. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 892–897.

[23] Qianxin Qu, Yijin Xiong, Xinyu Zhang, Chen Xia, Qian Peng, Ziqiang Song, Kang Liu, Xin Wu, and Jun Li. 2024. V2I-Calib++: A Multi-terminal Spatial Calibration Approach in Urban Intersections for Collaborative Perception. *arXiv preprint arXiv:2410.11008* (2024).

[24] Shunli Ren, Siheng Chen, and Wenjun Zhang. 2022. Collaborative perception for autonomous driving: Current status and future trend. In *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*. Springer, 682–692.

[25] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 2020. 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*. Springer, 605–621.

[26] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. 2021. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 1155–1162.

[27] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. 2023. 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*. 13712–13722.

[28] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. 2022. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*. Springer, 107–124.

[29] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. 2022. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2583–2589.

[30] Yan Yan, Yuxing Mao, and Bo Li. 2018. Second: Sparsely embedded convolutional detection. *Sensors* 18, 10 (2018), 3337.

[31] Dingkang Yang, Kun Yang, Yuzheng Wang, Jing Liu, Zhi Xu, Rongbin Yin, Peng Zhai, and Lihua Zhang. 2023. How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception. *Advances in Neural Information Processing Systems* 36 (2023), 25151–25164.

[32] Junwei You, Haotian Shi, Zhuoyu Jiang, Zilin Huang, Rui Gan, Keshu Wu, Xi Cheng, Xiaopeng Li, and Bin Ran. 2024. V2x-vlm: End-to-end v2x cooperative autonomous driving through large vision-language models. *arXiv preprint arXiv:2408.09251* (2024).

[33] Duanrui Yu, Jing You, Xin Pei, Anqi Qu, Dingyu Wang, and Shaocheng Jia. 2025. Which2comm: An Efficient Collaborative Perception Framework for 3D Object Detection. *arXiv preprint arXiv:2503.17175* (2025).

[34] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. 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*. 21361–21370.

[35] Yunshuang Yuan, Hao Cheng, and Monika Sester. 2022. Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3054–3061.

[36] Xinyu Zhang, Yijin Xiong, Qianxin Qu, Renjie Wang, Xin Gao, Jing Liu, Shichun Guo, and Jun Li. 2024. Cooperative Visual-LiDAR Extrinsic Calibration Technology for Intersection Vehicle-Infrastructure: A review. *arXiv preprint arXiv:2405.10132* (2024).

[37] Yin Zhou and Oncel Tuzel. 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4490–4499.