CoSTr: a Fully Sparse Transformer with Mutual Information for Pragmatic Collaborative Perception

Anonymous Authors

***Abstract*— Collaborative perception faces critical pragmatic challenges under bandwidth constraints, particularly in guaranteeing perception performance while overcoming spatial-temporal imperfections such as sensor misalignments and delays. Although recent fully sparse approaches have improved computational efficiency, they remain constrained by local receptive fields that limit long-range reasoning, lacking intelligent communication redundancy reduction, and still exhibit limited robustness to real-world noise. To address these challenges, we propose CoSTr, a Collaborative Sparse Transformer framework that operates natively on sparse feature representations through integrated sparse convolution and transformer. Our approach introduces a robust spatial-temporal attention mechanism that explicitly compensates for pose errors and communication delays during feature fusion. To minimize communication overhead, we propose a mutual information-based criterion that operates directly on sparse features to select and transmit only the most critical information. Extensive experiments on OPV2V, V2XSet and DAIR-V2X demonstrate that CoSTr not only achieves state-of-the-art perception performance and communication efficiency, but also significantly improves the robustness against spatial-temporal disturbances, establishing a new benchmark for pragmatic collaborative perception.**

I. 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 in the Intelligent Transportation System (ITS) to exchange their perceptual information and fuse multi-view observations from surrounding traffic participants, thereby constructing a more comprehensive, wide-area, and robust environmental representation that ensures safe transportation [1], [2], [3], [4], [5], [6], [7]. Following these works, we focus on LiDAR-based collaborative perception. Such collaboration mitigates blind spots and partial occlusions inherent in single-vehicle perception system, providing a stronger foundation for ITS to make decisions and plan trajectories in complex scenarios.

Nevertheless, as the perceptual range expands due to collaborative perception tasks, the computational cost of processing dense bird's-eye view (BEV) features grows quadratically. Moreover, under the communication bandwidth constraints typical in V2X scenarios, redundancy removal in long-range dense BEV features is not only computationally complex but, more critically, often leads to suboptimal collaborative perception performance. Therefore, to ensure both optimal perception performance and consistent computational overhead across varying perceptual ranges, sparse networks have garnered increasing attention and been applied to the field of collaborative perception. Sparse networks exhibit linear computational complexity relative to the number of non-empty points [8], and their output feature volume is significantly smaller than that of dense representations, making it easier to meet the communication constraint of 27MB/s bandwidth limitation specified by V2X communication standards [9], [10]. However, as illustrated in Figure 1, existing approaches that solely rely on sparse convolution for collaborative perception are fundamentally constrained by their **local receptive field** and spatial-temporal alignment are not tackled on sparse feature level. More importantly, while the volume of sparse features are naturally smaller than that of dense ones, the collaboration without **information exchanging** or **semantical redundancy removal** may still result in inefficient communication. Without a smart, learned selection mechanism, many transmitted features might contribute little to the collaborative perceptual task.

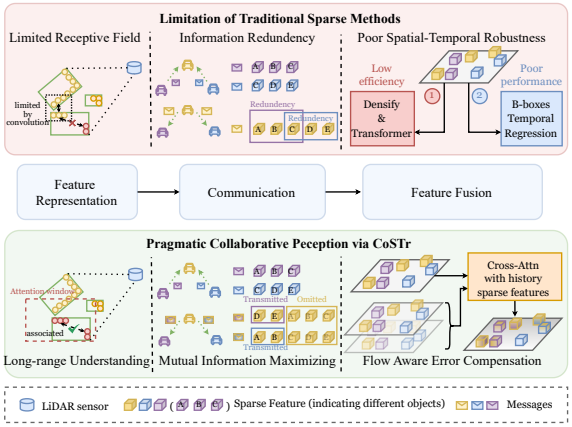


Fig. 1. Conceptual comparison of sparse collaborative perception paradigms. Compared with previous methods, the proposed CoSTr provides a pragmatic solution: It captures long-range dependencies to reason about consecutive areas, selectively exchanges only the informative features based on sparse mutual information, and achieves robust fusion under real-world constraints, enabling accurate and reliable collaborative perception.

To overcome the above challenges, we propose a novel collaborative perception framework — **CoSTr**, a **Co**llaborative **S**parse **Tr**ansformer that achieves **pragmatic collaborative perception** with both **spatial-temporal robustness** and **communication efficiency**. For spatial-

temporal robustness, CoSTr incorporate sparse transformer that enhances the sparse BEV features by capturing long-range contextual information through a sparse self-attention mechanism, and that effectively tackles communication delays and positioning errors during collaboration through cross-attention over sparse BEV flow, all these operations are processed inside partitioned window for computation efficiency. For communication efficiency, we introduce a sparse mutual information formulation based on [11] that selects the most informative features for transmission. This formulation operates directly on sparse feature sets, avoiding computationally expensive dense mutual information estimation, and combines mutual information maximization with a sparsity-inducing constraint to ensure minimal communication cost.

Our main contributions are summarized as follows:

- We propose a novel collaborative perception framework that fully operates on sparse features to achieve both spatial-temporal robustness and communication efficiency, advancing the pragmatic LiDAR-based collaborative perception under ITS scenario.
- We develop window-based sparse transformer to effectively enable long-range comprehension and compensate for delays and pose errors across collaborating agents through sparse attention mechanism, ensuring spatial-temporal robust collaborative perception.
- We propose a sparse mutual information formulation that operates directly on sparse features, enabling efficient selection of informative features for transmission under object-level bandwidth constraints, ensureing communication-efficient collaborative perception.
- We demonstrate through experiments that our approach effectively mitigates the limitations of existing sparse collaborative methods, offering improved performance and scalability in complex collaborative perce scenarios.

II. RELATED WORKS

A. Data Fusion in Collaborative Perception

Collaborative perception improves perception performance by sharing information among multiple agents, and can be divided into three categories according to the stage of data fusion: early fusion, intermediate fusion, and late fusion. **Early fusion** in DiscoNet [12] retains the original modality and interpretability, preserving full visual and 3D information. **Late fusion** transmits perception results for collaboration, achieving efficient communication but leading to weaker collaboration performance and robustness. Based on collaborative perception datasets such as [13], [14], [15], mainstream algorithms tend to adopt **intermediate fusion** approaches to achieve a better performance-communication trade-off [16], [17], [18], [19], [20], [21], [22], [23]. When2comm [24] introduces a handshake mechanism to select communication agents, V2X-ViT [25] designs a unified transformer to capture inter-agent interaction and spatial relationship, Where2comm [16] transmits features to specific locations based on spatial confidence, How2comm [17] and FF-Net [26] further incorporates a temporal prediction model. Pragcomm [27] alternates between transmitting local dynamic and global features, CodeFilling [18] maintains a codebook among agents and transmits integer codes. Which2comm [28] obtains sparse objects features through sparse convolution network but achieves temporal compensation with dense temporal attention mechanism, while SparseAlign [29] aligns spatial-temporal errors through object-level re-registration.

However, while these methods predominantly rely on dense BEV feature maps for processing and communication—delivering strong perceptual performance—they incur substantial communication and computational costs, particularly in multi-agent and complex scenarios. Sparse convolution-based methods turn to dense transformer [28] or object-level alignment [29] for help when it comes to spatial-temporal error robustness, resulting in either extra computation redundancy or suboptimal perception results.

B. Sparse Network for Object Detection

Sparse networks have emerged as a pivotal architecture for 3D object detection, effectively leveraging the inherent sparsity of point cloud data to achieve remarkable computational efficiency. Unlike dense networks that process entire feature grids [16], [17], [20], [30], [31], sparse approaches—including but not limited to sparse convolution—selectively operate only on non-empty regions [8], [32], [33], drastically reducing both computational overhead. VoxelNeXt [34] streamlined detection by adopting a fully sparse convolution pipeline, eliminating dense components even in the detection head. Similarly, the fully sparse detector in [35] retains and enhances fine-grained point-level features throughout the detection process. Which2Comm [28], SparseAlign [29] both adopt sparse convolutions for sparse collaborative perception. Beyond convolution, CoBEVT [21] apply sparse attention on dense features for high efficiency, while SST [36], FSTR [33] and FSHNet [37] introduce sparse transformer to capture long-range interactions within sparse features through carefully designed sparse attention mechanisms. These methods collectively establish that sparse networks gradually become a foundational design choice that maintains competitive accuracy while enabling scalable processing of large-scale 3D scenes. Notably, the integration of sparse transformers for rectifying spatiotemporal inconsistencies across agents could substantially improve the robustness of collaborative perception systems.

III. METHODOLOGY

To fully utilize the efficiency on both communication and computation of the sparse network, and to further avoid the redundancy on information exchanging and drawback on global, dynamic scenario comprehension, CoSTr introduces mutual information supervision for the sparse feature communication, and integrates sparse transformer into the sparse convolution structures. Figure 2 illustrate the overall architecture of CoSTr. Specifically, connected autonomous vehicles (CAVs) operate on sparse feature representations through sparse BEV extractor and Density-Aware Sparse Transformer (DAST) to generate the original information flow for collaboration, which will be further pruned under

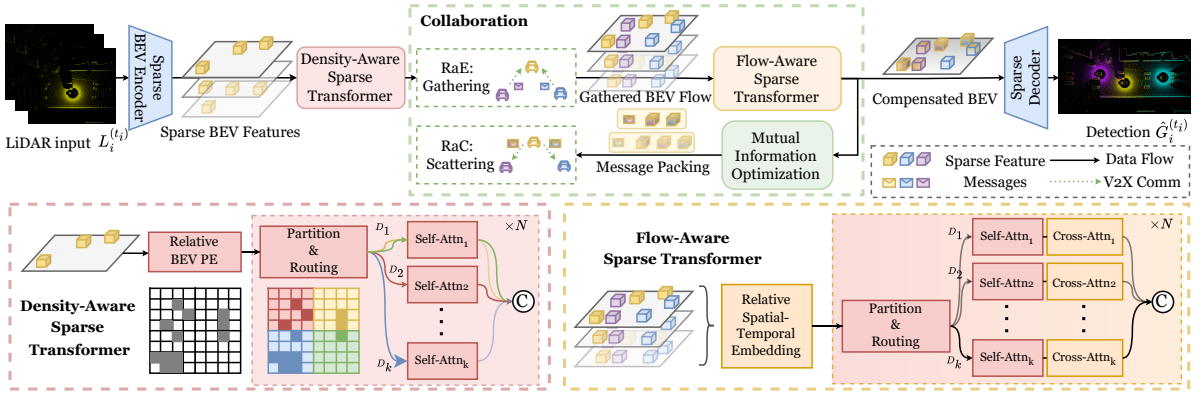


Fig. 2. Overall architecture of the proposed **CoSTr** framework with detailed structure of **DAST** and **FAST**. During Collaboration, CAV_i first serves as **RaE** to gather features for **FAST** and perception, then serves as **RaC** to reply optimized packed messages to collaborating CAVs.

the guidance of Mutual Information-based Communicating Optimization for efficient communication. Sparse features from different CAVs are gathered and the feature flow will be processed through Flow-Aware Sparse Transformer (FAST) that deals with the communication delay and positioning error, and finally the sparse detector decodes the compensated BEV to achieve the collaborative perception results.

A. Problem Formulation

Consider a system with K CAVs. Let $\mathbf{L}_i^{(t_i)}$ denote the point cloud captured by CAV_i at timestamp t_i, and $\mathbf{G}_i^{(t_i)}$ represent the corresponding ground truth for collaborative perception. In the collaborative perception pipeline as shown in Fig 2, The workflow for CAV_i can be formalized as:

$$\mathbf{F}_i^{(t_i)} = \Phi_{\text{enc}}(\mathbf{L}_i^{(t_i)}), \tag{1a}$$

$$\mathbf{M}_{i\to j}^{(t_i)} = \Phi_{\text{comp}}(\mathbf{F}_i^{(t_i)}, \mathbf{M}_{j\to i}^{(t_j)}), \tag{1b}$$

$$\hat{\mathbf{G}}_i^{(t_i)} = \Phi_{\text{dec}}\left(\Phi_{\text{fuse}}\left(\mathbf{F}_i^{(t_i)}, \{\mathbf{M}_{j\to i}^{(t_j)}\}_{j\neq i}^{K}\right)\right). \tag{1c}$$

where $\mathbf{F}_i^{(t_i)}$ denotes encoded features, $\mathbf{M}_{j\to i}^{(t_j)}$ represents compressed collaborative messages from the CAV_j with its corresponding timestamp $t_j < t_i$, reflecting the pragmatic communication delay; $\Phi_{\text{enc}}(\cdot)$,$\Phi_{\text{comp}}(\cdot)$, $\Phi_{\text{fuse}}(\cdot)$ 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 optimize performance under communication constraints, a single vehicle gathers and scatters collaborative information through the following dual-role mechanism:

- **Role as the Ego Vehicle (RaE)**: In a collaboration round, CAV_i first acts as the ego vehicle to aggregate relevant information from collaborators and fuses it with local observations following Equation (1c).
- **Role as a Collaborative Vehicle (RaC)**: After utilizing information from other vehicles, CAV_i turns into a collaborative vehicle to serve others, generating compact collaborative messages following Equation (1b).

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

$$\underset{\Phi_{\text{enc}},\Phi_{\text{comp}},\Phi_{\text{fuse}},\Phi_{\text{dec}}}{\arg\max} \sum_{i=1}^{K} \mathcal{E}_i \quad \text{s.t.} \quad \sum_{j\neq i} \mathcal{V}(\mathbf{M}_{j\to i}) \leq \mathbf{B}, \tag{2}$$

where $\mathcal{E}_i = g\left(\hat{\mathbf{G}}_i^{(t_i)}, \mathbf{G}_i^{(t_i)}\right)$, with $g(\cdot,\cdot)$ measuring the performance of perception of the CAV_i to be $\mathcal{E}_i$, and $\mathcal{V}(\cdot)$ calculates communication volume of messages.

B. Sparse BEV Encoder

After processing the point cloud $\mathbf{L}_i^{t_i}$ into pillars, the sparse BEV encoder operates on the encoded 2D pillar features into sparse BEV features $\{\mathbf{F}_i^{1\times}, \mathbf{F}_i^{2\times}, \mathbf{F}_i^{4\times}, \mathbf{F}_i^{8\times}\}$. Each CNN block is substituted with a 2D sparse convolution block. And in each layer of the sparse convolution block, focal loss [38] supervision is adopted to dynamically prune the features located outside the ground truth bounding-boxes, forcing sparse features neighboring to the boxes. Additionally, 2D additional down-sampling is applied similar to the one mentioned in [34] to obtain $\{\mathbf{F}_i^{16\times}, \mathbf{F}_i^{32\times}\}$, which ensures the effective receptive fields (ERFs) to be larger, thus enabling more accurate perception. The results of the ablation studies in Section IV.D demonstrate that the sparse feature extractor is an effective strategy to enhance performance. The overall sparse BEV extractor serves as the feature encoder mentioned in Equation (1a), where the sparse BEV features $\mathbf{F}_i^{bev} \in \mathbb{R}^{n_f \times c_f}$ and their corresponding indices and $\mathbf{I}_i^{bev} \in \mathbb{R}^{n_f \times 3}$ are extracted as follow:

$$\begin{aligned}
\mathbf{F}_i^{bev} &= \mathbf{F}_i^{8\times} \cup (\mathbf{F}_i^{16\times} \cup \mathbf{F}_i^{32\times}), \\
\mathbf{I}_i'^{32\times} &= \left\{(x_p \times 2^2,\ y_p \times 2^2) \mid p \in \mathbf{I}_i{}^{32\times}\right\}, \\
\mathbf{I}_i'^{16\times} &= \left\{(x_p \times 2^1,\ y_p \times 2^1) \mid p \in \mathbf{I}_i{}^{16\times}\right\}, \\
\mathbf{I}_i^{bev} &= \mathbf{I}_i{}^{8\times} \cup (\mathbf{I}_i'^{16\times} \cup \mathbf{I}_i'^{32\times}).
\end{aligned} \tag{3}$$

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

C. Density-Aware Sparse Transformer

The Density-Aware Sparse Transformer (DAST) module is designed to enhance the representational capacity of sparse

BEV features by efficiently capturing long-range contextual information through window-based sparse self-attention before cross-agent collaboration. As proved in SwinTransformer [39] that applying attention on partitioned windows leads to efficient computation, we further promote the efficiency to sparse features through density-aware routing. Building upon the sparse BEV features $(\mathbf{F}_i^{bev}, \mathbf{I}_i^{bev})$, the DAST module consists of three core components: Relative BEV Position Encoding, Attention Window Partition & Density-Aware Routing, and Sparse Self-Attention.

1) Relative BEV Position Encoding: To maintain spatial awareness within each attention window and ensure training stability, we introduce a relative BEV position encoding with a period equal to the window size. For a certain feature located at (x, y) within the window of size $S \times S$, the relative BEV position encoding is defined as:

$$\mathrm{PE}_{\text{rel}}(x, y) = \mathrm{PE}_{\text{bev}}\left(x \bmod S,\ y \bmod S\right), \tag{4}$$

while the encoded features are then augmented as $\mathbf{F}_i^{\text{enc}} = \mathbf{F}_i^{bev} + \mathrm{PE}_{\text{rel}}(\mathbf{I}_i^{bev})$. This periodic encoding scheme preserves spatial relationships within each window while ensuring consistency across different window locations.

2) Partition & Routing & Attention: The sparse BEV feature map is partitioned into non-overlapping windows of size $S \times S$. To handle varying feature density across windows and avoid unnecessary computation on empty regions, we employ a density-aware routing mechanism. Only windows containing features are processed, significantly improving computational efficiency. For each non-empty window w, the features are gathered as $\mathbf{F}_w = \{\mathbf{f}_p \mid p \in w \cap \mathrm{supp}(\mathbf{F}_i^{\text{enc}})\}$, where $\mathrm{supp}(\mathbf{F}_i)$ denotes the support of non-empty features. Though every window covers a considerably large region, there are some objects inevitably truncated. To tackle the truncation and enable cross-window information exchange, we employ a shifted window strategy where the partition pattern is shifted by $(S/2, S/2)$ in alternating layers, which allows the model to capture cross-window dependencies while maintaining computational efficiency.

Density-aware routing mechanism that dynamically adjusts the computational pathway based on the number of features within each window. Specifically, windows are categorized into different density intervals according to their feature count n_w. Each specific density level is associated with a corresponding self-attention module.

The routing process can be formalized as:

$$\mathrm{Route}(w) = \begin{cases} \mathrm{Attn}_{\text{Low}}(w) & \text{if } n_w/S^2 \le T_1, \\ \mathrm{Attn}_{\text{Medium}}(w) & \text{if } T_1 < n_w/S^2 \le T_2, \\ \mathrm{Attn}_{\text{High}}(w) & \text{if } n_w/S^2 > T_2, \end{cases} \tag{5}$$

where T_1 and T_2 are threshold parameters defining the density intervals, and $\mathrm{Attn}_*(\cdot)$ represents the attention module specialized for the corresponding density range. Within each window, we apply windowed multi-head self-attention (WMSA) [39]. The output features are computed through:

$$\tilde{\mathbf{F}}_w = \mathrm{WMSA}(\mathbf{F}_w) + \mathbf{F}_w, \tag{6}$$

$$\mathbf{F}_w^{\text{out}} = \mathrm{FFN}(\mathrm{LayerNorm}(\tilde{\mathbf{F}}_w)) + \tilde{\mathbf{F}}_w, \tag{7}$$

D. Flow-Aware Sparse Transformer

On the basis of the DAST, Flow-aware Sparse Transformer operates on the gathered BEV features from multiple CAVs and further incorporates mechanisms for temporal feature propagation and cross-agent spatial alignment, achieving robust feature fusion under real-world constraints including communication delays and positioning errors.

1) Inputs and Feature Gathering: The CAV_i accepts as input the sparse BEV features $\{\mathbf{F}_j^{(t_j)}\}_{j\neq i}^{N}$ and their corresponding indices $\{\mathbf{I}_j\}_{j\neq i}^{N}$ at time t_j from collaborative agents. These features and indices are gathered into a unified sparse tensor through the concatenation operation, together with history BEV features, forming the Gathered BEV Flow.

2) Relative Spatial-Temporal Embedding: To achieve effective spatio-temporal alignment of features from collaborative agents captured at different times and under different perspectives, we propose a unified spatio-temporal positional encoding. Let $\mathbf{I}_i^{(t_0)}$ denote the BEV indices of the ego agent at the current timestamp t_0, and $\mathbf{I}_j^{(t_j)}$ represent the BEV indices from the j-th collaborative agent captured at time t_j. The combined encoding is formulated as:

$$\begin{aligned} &\mathrm{PE}_{rst}(\mathbf{I}_i^{(t_0)}, \mathbf{I}_j^{(t_j)}, t_0, t_j, \mathbf{T}_j^{(t_j)}) \\ =&\mathrm{PE}_{\text{rel}}\left((\mathbf{I}_i^{(t_0)}, \mathbf{T}_j^{(t_j)}\mathbf{I}_j^{(t_j)})\right) + \mathrm{PE}_{\text{temp}}(t_0 - t_j), \end{aligned} \tag{8}$$

where PE_{rel} denotes the relative BEV PE in Equation (4), $\mathbf{T}_j^{(t_0)}\mathbf{I}_j^{(t_j)}$ transforms the collaborative indices into the local coordinate of the ego vehicle, and $\mathrm{PE}_{\text{temp}}$ is the temporal embedding defined as:

$$\mathrm{PE}_{\text{temp}}(\Delta t)[k] = \begin{cases} \sin(\omega_k \Delta t) & \text{if } k \text{ is even}, \\ \cos(\omega_k \Delta t) & \text{if } k \text{ is odd}, \end{cases} \tag{9}$$

with $\omega_k = 1/(10000^{2k/d})$, d being the embedding dimension, and Δt the time delay from received timestamp t_j to current timestamp t_0. Then features can be encoded as $\mathbf{F}_{fuse}^{(t_0)} = (\mathbf{F}_i^{(t_0)}, \mathbf{F}_j^{(t_j)}) + \mathrm{PE}_{rst}(\mathbf{I}_i^{(t_0)}, \mathbf{I}_j^{(t_j)}, t_0, t_j, \mathbf{T}_j^{(t_j)})$. This enhanced feature representation encapsulates both spatial context and temporal state.

3) Window-Based Temporal Cross-Attention: The gathered BEV features are processed through the structure same as that in DAST until the cross attention mechanism inside each window between current features $\mathbf{F}_{fuse}^{(t_0)}$ and temporal embedded history features $(\mathbf{F}_{\text{comp}}^{(t_{\text{hist}})} + \mathrm{PE}_{\text{temp}}(t_0 - t_{hist}))$ as $\mathbf{F}_{\text{comp}}^{(t_0)} = \text{Cross-Attn}\left(\mathbf{F}_{fuse}^{(t_0)}, (\mathbf{F}_{\text{comp}}^{(t_{\text{hist}})} + \mathrm{PE}_{\text{temp}}(t_0 - t_{hist}))\right)$. This design effectively compensates for spatial misalignment and communication delays, ensuring robust collaborative perception even under challenging V2X conditions.

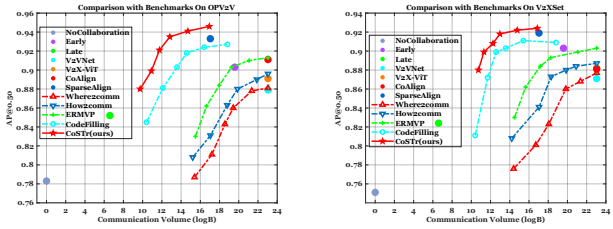
E. Sparse Feature Selection

Effectively selecting and transmitting the most informative features under extreme bandwidth constraints is a central challenge in collaborative perception. To address this, we introduce a sparse mutual information (MI) criterion. MI provides a theoretically-grounded measure of how much information the feature exclusively conveys, which is critical

TABLE I

PERFORMANCE COMPARISON AND COMMUNICATION COST ON OPV2V [15], V2XSET [25] AND DAIR-V2X [13].

Methods	OPV2V			V2XSet			DAIR-V2X		
	AP@0.5↑	AP@0.7↑	AB ↓	AP@0.5↑	AP@0.7↑	AB↓	AP@0.5↑	AP@0.7↑	AB ↓
No Fusion	0.783	0.663	0	0.751	0.618	0	0.512	0.342	0
Early	0.903	0.797	19.59	0.884	0.771	19.81	0.55	0.385	19.09
Late	0.852	0.744	6.59	0.824	0.703	6.04	0.534	0.361	5.44
V2VNet [19]	0.879	0.780	23.04	0.871	0.785	23.04	0.561	0.446	22.01
Where2comm [16]	0.881	0.793	23.04	0.877	0.784	23.04	0.553	0.441	22.01
V2X-ViT [25]	0.891	0.798	23.04	0.882	0.789	23.04	0.558	0.446	22.01
How2comm [17]	0.896	0.803	23.04	0.887	0.794	23.04	0.562	0.451	22.01
ERMVP [40]	0.913	0.842	23.04	0.903	0.828	23.04	0.573	0.463	22.01
CodeFilling [18]	0.927	0.896	18.81	0.909	0.840	18.81	0.609	0.505	17.74
CoAlign [41]	0.911	0.874	23.04	0.881	0.835	23.04	0.586	0.492	22.01
SparseAlign [29]	0.933	0.901	17.04	0.919	0.847	17.01	0.613	0.511	16.95
CoSTr (ours)	**0.946**	**0.916**	**16.94**	**0.924**	**0.859**	**16.87**	**0.621**	**0.522**	**15.23**



(a) Comparison on OPV2V (b) Comparison on V2XSet

Fig. 3. Comparison with benchmarks of the trade-off between perception performance and communication volume on OPV2V [15] and V2XSet [25], CoSTr retains high performance under object-level communication limitations, achieving the best perception-communication trade-off.

for ensuring the collaborative perception performance while maintaining minimal communication cost. The sparse formulation of MI in this section avoids computationally expensive dense MI estimation and operates directly on sparse features with spatial indices.

1) Window-based Feature Aggregation: To efficiently compute mutual information between large-scale sparse feature sets $(\mathbf{X}, \mathbf{I_X})$ and $(\mathbf{Y}, \mathbf{I_Y})$, we partition the 2D space into $S \times S$ non-overlapping windows $\mathcal{W} = \{w_{mn}\}$. For each non-empty window, we aggregate features from both sets:

$$\mathbf{z}_{mn} = \text{avg}\left(\{\mathbf{f}_p^{\mathbf{X}}\}_{p \in \mathbf{I}_{mn}^{\mathbf{X}}} \cup \{\mathbf{f}_p^{\mathbf{Y}}\}_{p \in \mathbf{I}_{mn}^{\mathbf{Y}}}\right), \tag{10}$$

where $\mathbf{I}_{ij}^{\mathbf{X}}$ and $\mathbf{I}_{ij}^{\mathbf{Y}}$ denote Indices from $\mathbf{X}$ and $\mathbf{Y}$ falling within window w_{ij}. This aggregation preserves spatial coherence while significantly reducing computational complexity.

2) Window-level Mutual Information Estimation: We estimate mutual information using the InfoNCE lower bound, which provides a scalable and differentiable approximation:

$$\mathcal{I}(\mathbf{X}; \mathbf{Y}) = \frac{1}{K} \sum_{k=1}^{K} \left[\log \frac{\exp(\mathbf{z}_k^T \mathbf{z}_k / \tau)}{\sum_{m=1}^{M} \exp(\mathbf{z}_k^T \mathbf{z}_m / \tau)} \right], \tag{11}$$

where K is number of non-empty windows, and τ is a temperature parameter controlling the sharpness of distribution.

3) Sparse MI Optimization: The overall objective combines mutual information maximization with a sparsity-inducing constraint to ensure communication efficiency:

$$\mathcal{L}_{MI} = -\sum_{j \neq i} \mathcal{I}(\mathbf{F}_{\text{comp},i}^{(t_0)} \odot \mathbf{m}; \mathbf{F}_{\text{comp},j}^{(t_0)}) + \lambda \|\mathbf{m}\|_1, \tag{12}$$

where $\mathbf{m}$ is a learnable mask vector for feature selection, and λ controls the supervision strength. Thus, the communication volume from CAV_i to CAV_j can be calculated as:

$$\mathcal{V}(\mathbf{F}_{\text{comp},i}^{(t_0)} \odot \mathbf{m}) = len(\mathbf{m}) \times (c^f + 2) \times 4\text{B}, \tag{13}$$

where the factor $c^f + 2$ represents the number of output channels with a 2D spatial indices, while the factor 4B represents the byte-width of the *float32* data type.

IV. EXPERIMENT RESULT

A. Datasets and Implementation Details

Datasets. To evaluate the effectiveness of CoSTr on the collaborative perception task, experiments are conducted on three widely used multi-agent collaborative perception datasets. **OPV2V** [15] is a large-scale V2V dataset simulated on CARLA [42], containing over 70 driving scenes and 10,914 annotated LiDAR point cloud frames. **V2XSet** [25] is a simulated V2X dataset comprising 11,447 frames in total and includes up to 5 connected agents per scene. **DAIR-V2X** [13] is a real-world dataset containing 100 driving scenarios and 18,000 data samples from a vehicle and the infrastructure, and we adopt the original DAIR-V2X 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 objective comparison of various perception methods, the singular LiDAR point cloud modality is utilized as the input for all methods. The LiDAR range for CAVs in our experiment is $281.6m \times 76.8m$ on OPV2V and V2XSet, and $102.4m \times 102.4m$ on DAIR-V2X for the intersection scenario. The tested models share LiDAR backbone of PointPillars [43] with the default pillar resolution of $0.4m \times 0.4m$. Perception performance is evaluated with average precision (AP) at Intersection-over-Union (IoU) threshold of 0.5 and 0.7. The communication volume is evaluated by applying logarithmic operation with base 2 on the total bytes of the transmitting message, denoted as

TABLE II

ROBUSTNESS ANALYSIS ON OPV2V AND V2XSET DATASET UNDER DIFFERENT NOISE CONDITIONS (AP@0.5).

Dataset	Methods	Perfect	Delay d(s)				Heading $\sigma_h(^\circ)$				Position σ_p(m)			
			0.1	0.2	0.3	0.4	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
OPV2V	V2X-ViT	0.891	0.881	0.869	0.854	0.833	0.884	0.874	0.857	0.821	0.881	0.871	0.849	0.825
	How2Comm	0.896	0.892	0.883	0.859	0.844	0.883	0.875	0.853	0.831	0.890	0.878	0.865	0.837
	CoAlign	0.911	0.901	0.885	0.867	0.850	0.902	0.888	0.862	0.854	0.902	0.890	0.865	0.857
	SparseAlign	0.933	0.926	0.916	0.903	0.896	0.917	0.904	0.889	0.877	0.917	0.909	0.891	0.871
	CoSTr(ours)	**0.946**	**0.937**	**0.928**	**0.917**	**0.909**	**0.937**	**0.921**	**0.904**	**0.892**	**0.941**	**0.926**	**0.913**	**0.902**
V2XSet	V2X-ViT	0.882	0.874	0.863	0.849	0.820	0.871	0.865	0.844	0.827	0.872	0.868	0.845	0.826
	How2Comm	0.887	0.880	0.869	0.854	0.828	0.884	0.868	0.860	0.826	0.877	0.871	0.853	0.832
	CoAlign	0.881	0.874	0.862	0.848	0.84	0.875	0.858	0.851	0.834	0.875	0.859	0.853	0.838
	SparseAlign	0.919	0.915	0.907	0.894	0.880	0.913	**0.912**	0.887	0.884	0.911	0.910	0.892	0.885
	CoSTr(ours)	**0.924**	**0.920**	**0.914**	**0.902**	**0.893**	**0.915**	0.906	**0.904**	**0.890**	**0.919**	**0.913**	0.**906**	**0.895**

TABLE III

ABLATION STUDY RESULTS ON OPV2V DATASET. ABBREVIATION OF MODULES ARE USED: SE (SPARSE BEV ENCODER), DAST, FAST AND MI (SPARSE MUTUAL INFORMATION OPTIMIZATION).

SE	DAST	FAST	MI	AP@0.5*↑			AB↓
				Perf	D=0.2s	P=0.4m	
-	-	-	-	0.875	0.831	0.762	16.43
✓	-	-	-	0.897	0.842	0.79	16.94
✓	✓	-	-	0.919	0.849	0.883	16.94
✓	✓	DAST†	-	0.933	0.878	0.911	16.94
✓	✓	✓	-	**0.946**	**0.928**	**0.926**	16.94
✓	✓	✓	✓	0.941	0.924	0.925	**14.71**

$*$: AP@0.5 evaluated under different noise settings of perfect condition, $\sigma_d = 0.2s$, and $\sigma_p = 0.4m$.
†: substitute FAST with DAST module during collaboration.

"AB" column in Tables. All models are trained on NVIDIA RTX 3090 GPUs with the Adam optimizer with a learning rate of $2 \times e^{-4}$ for 80 epoches and 10 extra epoches for noisy settings. For attention windows, S is set to 8, and T_1, T_2 are set to 0.125 and 0.25, respectively. For FAST module, two history frames are loaded for temporal compensation. For robustness evaluation, we inject Gaussian-distributed spatial errors following [15] and simulate temporal asynchrony by introducing one frame mismatch for every 100 ms delay.

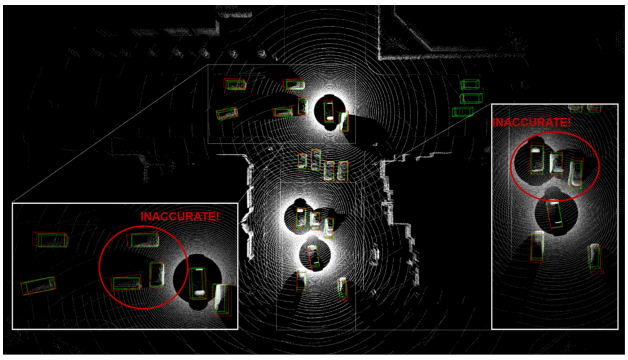
B. Quantitative Evaluation

1) Collaborative Perception Performance: Table I compares CoSTr with different intermediate fusion benchmarks including V2VNet [19], Where2comm [16], V2X-ViT [25], How2comm [17], ERMVP [40], CodeFilling [18], CoAlign [41] and SparseAlign [29]. Looking into the the comparison on the perception performance, our method outperforms dense SOTA CodeFilling by an improvement on AP of 1.5% at the IoU of 0.7 and 1.9% at the IoU of 0.5 on OPV2V mainly due to the inherent ability of sparse networks on filtering non-significant features and the enhanced ability on temporal reasoning. Comparing with sparse SOTA SparseAlign, CoSTr achieves higher performance on all datasets thanks to long-range contextual sensing ability that associates distant queries. Similar advantageous performance is consistently observed on the V2XSet and DAIR-V2X datasets.

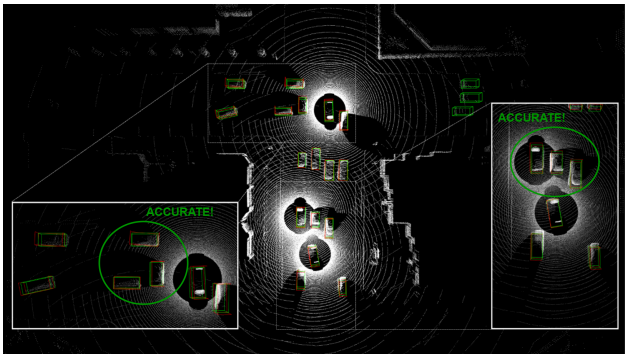
2) Communication Efficiency Comparison: As shown in Table I, while achieving the highest performance on all datasets, the logarithmic communication volume retains at rather low level. Furthermore, Figure 3 illustrates the comprehensive communication-perception trade-off landscape on the benchmark datasets. Comparing to previous SOTA SparseAlign, the communication volume of CoSTr can be controlled with the Mutual Information Optimization mechanism, while the curve corresponding to our CoSTr resides at the top-left corner of the graphs, indicating that it consistently delivers superior perception performance (higher AP) across the entire spectrum of communication volumes. Under strict object-level communication constraints, CoSTr retains high performance while most methods like ERMVP and How2Comm fail to achieve as same performance as late fusion. On the other hand, CoSTr surpass Codefilling on AP@0.5 with only $1/64\times$ communication volume on OPV2V. The results unequivocally validate that CoSTr sets a new state-of-the-art in pragmatic collaborative perception by optimally balancing the communication-perception trade-off.

3) Robustness Analysis: To evaluate robustness to pragmatic errors of communication asynchrony and position misalignment, we verify the performance of CoSTr under varying temporal asynchrony and spatial errors shown in Table II. For fair comparison, benchmarks in the analysis are all equiped with error rectifying ability. On OPV2V, under communication delays (d), baseline methods exhibit significant performance degradation of 4-6% at $d = 0.4s$, while CoSTr maintains stable performance with only a 3.7% decrease. Similarly, with heading errors (σ_h), CoSTr shows a minimal 5.4% performance reduction at $\sigma_h = 0.8^\circ$ compared to 6-8% degradation in other methods. For positioning errors (σ_p), CoSTr achieves the smallest performance decline (2.4% at $\sigma_p = 0.8m$) while other methods suffer 5-7% drops. Similar trend can be observed on V2XSet. This consistent robustness advantage stems from FAST module, which explicitly handles spatio-temporal inconsistencies through relative spatial-temporal embedding and cross-attention, ensuring reliable performance under pragmatic noises.

4) Computation Complexity: The sparse attention mechanism in DAST and FAST ensures that the computational resources are allocated proportionally to the actual information content of each window, avoiding unnecessary computations on near-empty windows while preserving modeling capacity for information-dense regions. The overall computational



(a) SparseAlign



(b) CoSTr

Fig. 4. Visualization of detection results from the OPV2V dataset. Green and red boxes represent the ground truth and detection boxes respectively.

complexity of DAST is bounded by:

$$\mathcal{O}(\text{DAST}) = NC^2 + S^2NC, \tag{14}$$

where N is the total number of non-empty features, C is the feature dimension, and S is the window size.

Compared to alternative approaches, our density-aware routing provides significant efficiency gains:

- Window Dense Attention: $\mathcal{O}(\text{Win-Dense}) = NC^2 + S^2HWC$ – suffers from redundant computation on empty indices, equivalent to dense WMSA.
- Global Sparse Attention: $\mathcal{O}(\text{Global Sparse}) = NC^2 + N^2C$ – becomes prohibitive for large N.

By adaptively distributing computational resources based on intra-window feature density, this method achieves the optimal balance between modeling capacity and computational efficiency, making it particularly suitable for the application on real-time and robust collaborative perception.

C. Ablation Studies

The ablation study in Table III systematically evaluates the contribution of each component in our framework.The incorporation of the sparse BEV encoder brings noticeable improvement in perception accuracy with marginal extra cost on communication volume. DAST further enhances performance significantly, particularly in handling complex spatial relationships and capturing long-range dependencies. When examining the collaboration mechanism, replacing FAST with the DAST module results in clear performance degradation, especially under challenging conditions with temporal delays and positioning errors. This underscores FAST's crucial role in robust spatio-temporal alignment during cross-agent fusion. The complete CoSTr framework without MI achieves the best perception performance, demonstrating the synergistic effect of integrating all components. Finally, the mutual information optimization (MI) module reduces communication overhead by $1/4.6\times$ with only marginal performance sacrifice, highlighting its effectiveness in identifying and transmitting the most valuable information.

D. Qualitative Evaluation

Figure 4 visualizes the detection results of SparseAlign and CoSTr in an intersection scenario from the OPV2V dataset under pragmatic noise setting of $d = 0.1s$, $\sigma_h = 0.2°$, $\sigma_p = 0.2m$. Qualitative comparisons clearly demonstrate CoSTr's superior perception capability. Our method produces detection boxes that align more precisely with ground truth annotations, exhibiting tighter bounding box fits and more accurate orientation estimation. These qualitative results corroborate the quantitative findings, confirming that our approach achieves more reliable and accurate perception under realistic noise conditions, validating the quantitative advantages demonstrated in our robustness analysis.

V. CONCLUSION

In this paper, we proposed **CoSTr**, a novel fully sparse collaboration framework that revisits communication-efficient collaborative perception through **sparse transformers** and **sparse mutual-information optimizaiton mechanism**. The former one enables SOTA performance and pragmatic robustness, while the latter one realizes excellent communication-performance trade-off. Evaluations on both simulated and real-world datasets validate that CoSTr significantly outperforms SOTAs across the communication-accuracy spectrum, while results under pragmatic noise setting also shows excellent robustness. Future research will investigate extending our sparse-centric methodology to heterogeneous sensor fusion scenarios, exploring cross-modal synergy between LiDAR, radar, and images. Such extensions could enable more robust environmental understanding through complementary modality-specific features, ultimately enhancing decision-making robustness for safety-critical ITS.

REFERENCES

[1] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets, and challenges," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 6, pp. 131–151, 2023.

[2] S. Ren, S. Chen, and W. Zhang, "Collaborative perception for autonomous driving: Current status and future trend," in *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*. Springer, 2022, pp. 682–692.

[3] A. Caillot, S. Ouerghi, P. Vasseur, R. Boutteau, and Y. Dupuis, "Survey on cooperative perception in an automotive context," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 204–14 223, 2022.

[4] Y. Yuan, H. Cheng, and M. Sester, "Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3054–3061, 2022.

[5] S. Liu, C. Gao, Y. Chen, X. Peng, X. Kong, K. Wang, R. Xu, W. Jiang, H. Xiang, J. Ma, *et al.*, "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," *arXiv preprint arXiv:2308.16714*, 2023.

[6] Q. Qu, Y. Xiong, G. Zhang, X. Wu, X. Gao, X. Gao, H. Li, S. Guo, and G. Zhang, "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, 2024, pp. 892–897.

[7] Q. Qu, Y. Xiong, X. Zhang, C. Xia, Q. Peng, Z. Song, K. Liu, X. Wu, and J. Li, "V2i-calib++: A multi-terminal spatial calibration approach in urban intersections for collaborative perception," *arXiv preprint arXiv:2410.11008*, 2024.

[8] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky, "Sparse convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 806–814.

[9] D. Jiang, Q. Chen, and L. Delgrossi, "Optimal data rate selection for vehicle safety communications," in *Proceedings of the fifth ACM international workshop on VehiculAr Inter-NETworking*, 2008, pp. 30–38.

[10] A. Filippi, K. Moerman, V. Martinez, A. Turley, O. Haran, and R. Toledano, "Ieee802. 11p ahead of lte-v2v for safety applications," *Autotalks NXP*, vol. 1, pp. 1–19, 2017.

[11] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," 2019. [Online]. Available: https://arxiv.org/abs/1808.06670

[12] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 541–29 552, 2021.

[13] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. 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*, 2022, pp. 21 361–21 370.

[14] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. 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*, 2023, pp. 13 712–13 722.

[15] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. 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)*. IEEE, 2022, pp. 2583–2589.

[16] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," *Advances in neural information processing systems*, vol. 35, pp. 4874–4886, 2022.

[17] D. Yang, K. Yang, Y. Wang, J. Liu, Z. Xu, R. Yin, P. Zhai, and L. Zhang, "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," *Advances in Neural Information Processing Systems*, vol. 36, pp. 25 151–25 164, 2023.

[18] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 481–15 490.

[19] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. 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*. Springer, 2020, pp. 605–621.

[20] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. 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*, 2019, pp. 88–100.

[21] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[22] R. Mao, H. Wu, Y. Jia, Z. Nan, Y. Sun, S. Zhou, D. Gündüz, and Z. Niu, "Diffcp: Ultra-low bit collaborative perception via diffusion model," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 6587–6593.

[23] Y. Xu, L. Li, J. Wang, B. Yang, Z. Wu, X. Chen, and J. Wang, "Codyntrust: Robust asynchronous collaborative perception via dynamic feature trust modulus," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 336–342.

[24] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 4106–4115.

[25] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European conference on computer vision*. Springer, 2022, pp. 107–124.

[26] H. Yu, Y. Tang, E. Xie, J. Mao, P. Luo, and Z. Nie, "Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[27] Y. Hu, X. Pang, X. Qin, Y. C. Eldar, S. Chen, P. Zhang, and W. Zhang, "Pragmatic communication in multi-agent collaborative perception," *arXiv preprint arXiv:2401.12694*, 2024.

[28] D. Yu, J. You, X. Pei, A. Qu, D. Wang, and S. Jia, "Which2comm: An efficient collaborative perception framework for 3d object detection," *arXiv preprint arXiv:2503.17175*, 2025.

[29] Y. Yuan, Y. Xia, D. Cremers, and M. Sester, "Sparsealign: A fully sparse framework for cooperative object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[30] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[31] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6876–6883.

[32] X. Lai, Y. Chen, F. Lu, J. Liu, and J. Jia, "Spherical transformer for lidar-based 3d recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[33] D. Zhang, Z. Zheng, H. Niu, X. Wang, and X. Liu, "Fully sparse transformer 3-d detector for lidar point cloud," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.

[34] Y. Chen, J. Liu, X. Zhang, X. Qi, and J. Jia, "Voxelnext: Fully sparse voxelnet for 3d object detection and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 674–21 683.

[35] L. Fan, F. Wang, N. Wang, and Z.-X. Zhang, "Fully sparse 3d object detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 351–363, 2022.

[36] L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, and Z. Zhang, "Embracing single stride 3d object detector with sparse transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8458–8468.

[37] S. Liu, M. Cui, B. Li, Q. Liang, T. Hong, K. Huang, Y. Shan, and K. Huang, "Fshnet: Fully sparse hybrid network for 3d object detection," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 8900–8909.

[38] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[40] J. Zhang, K. Yang, Y. Wang, H. Wang, P. Sun, and L. 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*, 2024, pp. 12 575–12 584.

[41] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3d object detection in presence of pose errors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 4812–4818.

[42] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[43] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.