2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Communication-Efficient and Context-Adaptive Collaborative Perception

Wenyu Lu[1], Hui Zhang*[1,2], Yuquan Yang[1], Ziyin Zhang[1], Xiaohua Xu[1]

***Abstract*— Collaborative perception is pivotal for the large-scale deployment of autonomous driving, yet it has long grappled with the trade-off between perception accuracy and bandwidth consumption. Existing methods fail to analyze the fine-grained characteristics of Field of View (FoV), leading to inefficient bandwidth utilization. To address this, we propose a Context-adaptive Collaborative Perception framework, termed CaCP. This method optimizes bandwidth usage by employing distinct collaboration strategies for FoV under varying contexts, thereby reducing communication overhead while maintaining perception accuracy. Additionally, CaCP introduces a novel spatial fusion of intermediate and late fusion strategies, yielding a more flexible collaborative scheme. Extensive experiments across multiple datasets encompassing both simulated (OPV2V) and real-world (V2V4Real) scenarios demonstrate that CaCP establishes a new state-of-the-art trade-off between accuracy and bandwidth. Notably, it reduces bandwidth consumption by up to 17% compared to previous works while achieving competitive or superior perception performance.**

I. Introduction

Accurate environmental perception is fundamental for intelligent agents (e.g., autonomous vehicles) to ensure safety in complex driving environments. However, single-agent perception systems face inherent limitations: occlusion and restricted detection range [1]. Multi-agent Collaborative perception via Vehicle-to-Everything (V2X) communication addresses these issues by integrating distributed viewpoints into a global perspective [2]. While enhancing perception, V2X introduces a critical trade-off: high perception accuracy demands substantial data transmission, conflicting with limited bandwidth resources, ultimately manifesting as a fundamental **accuracy-bandwidth trade-off**.

This trade-off has spurred a spectrum of solutions ranging from early fusion to late fusion. Early fusion methods [3], [4], [5] exchange raw sensor measurements (e.g., LiDAR point clouds, camera images) to minimize occlusion effects. While capable of achieving optimal performance, they incur prohibitively high communication bandwidth costs. Conversely, late fusion methods [6], [7] exchange high-level outputs (e.g., object detection results represented as 3D bounding boxes) to minimize bandwidth usage. However, studies show that the performance gains from this approach are often very limited [8]. To balance these competing design objectives, research in V2X collaborative frameworks has gravitated towards the middle ground of the spectrum—intermediate fusion [9], [8], [10]. This paradigm involves selectively broadcasting intermediate feature representations, such as Bird's-Eye View (BEV) maps of the agent's surrounding environment. Currently, intermediate fusion dominates research due to its better balance between accuracy and bandwidth consumption.

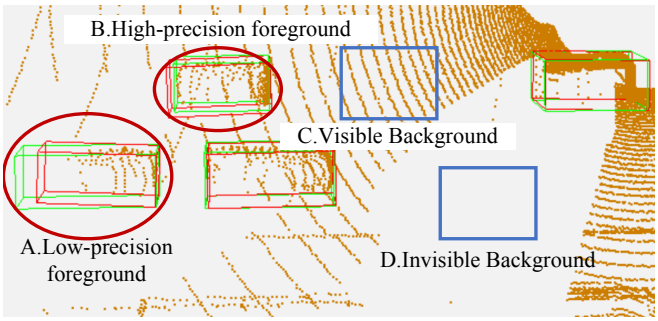

Fig. 1. Point cloud and FoV classification. Green 3D bounding boxes: ground truth. Red 3D bounding boxes: detections. Red ellipses: foreground regions (A/B). Blue rectangles: background regions (C/D).

Nevertheless, current intermediate fusion methods remain far from optimal. Collaborating using uncompressed intermediate features (e.g., BEV maps) can consume even more bandwidth than early fusion methods that directly share raw point clouds [7]. The most prevalent approach employs a compression module to reduce intermediate feature maps, thereby decreasing communication overhead. However, high compression ratios inevitably lead to significant accuracy degradation. We argue that the primary source of this inefficiency is the indiscriminate broadcasting paradigm, where a uniform communication strategy (e.g., feature compression) is applied across the entire FoV. Such context-agnostic approaches treat all regions as equally important, failing to distinguish between critical objects requiring detailed features and clear background areas. Consequently, bandwidth is squandered on transmitting redundant information (e.g., features of a high-precision, easily detectable vehicle) or low-value data (e.g., occluded areas). Researchers have recognized this issue and initiated preliminary explorations. For instance, Where2comm [11] attempts to reduce bandwidth by transmitting features only from detected foreground regions. More recently, CoSDH [12] designed a novel supply-and-demand-aware information selection module to enhance communication efficiency by optimizing the selection of

*Corresponding author.
[1]School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230027, China.
Email: (Wenyu Lu, Yuquan Yang, Ziyin Zhang){luwenyu9, yuquany, zhangziyin5163}@mail.ustc.edu.cn,
(Hui Zhang, Xiaohua Xu){fzhh, xiaohuaxu}@ustc.edu.cn.
[2]School of Computer Science and Technology, University of Science and Technology of China, Anhui Provincial Key Laboratory of High Performance Computing, Hefei, 230027, China.

required regions, while also incorporating late fusion to boost detection accuracy. However, these methods primarily focus on selecting which parts of the FoV to share, lacking finer-grained situational analysis and failing to adapt strategies based on distinct context types.

As illustrated in Fig. 1, our key insight is to replace this indiscriminate strategy with a context-adaptive one. We perform a fine-grained partitioning of the FoV into four distinct contexts, each with a tailored, bandwidth-efficient collaboration policy: (A) Low-Precision Foreground regions, where uncertainty is high, necessitate sharing rich intermediate features for collaborative disambiguation. (B) High-Precision Foreground regions, where the local agent's perception is already reliable, only require lightweight detection results (i.e., late fusion) for confirmation and redundancy removal. (C) Visible Background regions, containing valuable negative evidence, are efficiently encoded via a simple binary mask to suppress false positives. Finally, (D) Invisible Background regions are correctly identified as containing no useful information, thus consuming zero bandwidth. This systematic allocation ensures that bandwidth is prioritized for regions of highest informational value.

Building upon this analysis, we propose a context-adaptive collaborative perception (**CaCP**) framework, which intelligently selects information and fusion strategy per region context. This approach prioritizes allocating precious communication bandwidth to the regions with the highest uncertainty and the greatest value for collaborative perception, while still capturing other critical information effectively, thereby achieving more efficient bandwidth utilization. Crucially, this process inherently implements a spatial fusion of intermediate and late fusion strategies, resulting in a more flexible collaborative scheme. The CaCP framework comprises two main components: i) a Context-Aware Message Packing module that identifies the context of each region and selects information for transmission based on the collaboration strategy; and ii) an Information Fusion and Detection module that effectively integrates the received heterogeneous information to produce the final detection results.

In summary, our contributions are as follows:

- We propose a novel Context-Adaptive Collaborative Perception framework that achieves a new state-of-the-art accuracy-bandwidth trade-off.
- We introduce a systematic, fine-grained contextual partitioning of the FoV, which dynamically assigns a tailored communication policy (intermediate fusion, late fusion, or a compact mask) to each region based on its context.
- We conduct extensive experiments on both simulated and real-world datasets, demonstrating CaCP's superior performance and its unique ability to adaptively increase communication to ensure detection accuracy in challenging scenarios.

II. Related Work

A. Single-Agent 3D Object Detection

3D object detection serves as a fundamental component in autonomous driving systems, encompassing image-based [13], [14], point-cloud-based [15], [16], and multi-modal fusion approaches [17], [18]. While these methods have achieved remarkable progress, they inherently suffer from occlusion and limited sensing range when operating independently. This motivates the development of collaborative perception systems that leverage multi-agent information sharing to overcome single-agent limitations.

B. Collaborative Perception

Collaborative perception methods can be generally categorized into early fusion [3], [5], intermediate fusion [9], [10], and late fusion [6], [7]. Intermediate fusion has emerged as the dominant paradigm due to its superior balance between detection performance and communication efficiency.

Performance-Oriented Methods. Early intermediate fusion works focus primarily on improving detection accuracy. Representative methods include F-Cooper [19] with element-wise feature aggregation, V2VNet [9] leveraging graph neural networks, and DiscoNet [8] employing knowledge distillation. Recent transformer-based approaches like CoBEVT [20] and V2X-ViT [10] achieve state-of-the-art performance through sophisticated attention mechanisms. However, these methods typically transmit complete intermediate features, resulting in prohibitive bandwidth consumption for practical deployment.

Communication-Efficient Methods. To address bandwidth limitations, several approaches have been proposed to compress collaborative information. The fundamental strategy involves channel-wise compression using autoencoders [7], but suffers from significant performance degradation at high compression ratios. Where2comm [11] introduces spatial sparsity by transmitting only foreground regions, yet its confidence-based selection strategy lacks efficiency. Subsequent works like FPV-RCNN [21] and CodeFilling [22] apply advanced compression techniques but fail to optimize the fundamental information selection process. CoSDH [12] models supply-demand relationships between agents and incorporates late fusion as compensation. However, it remains constrained by a binary transmit-or-not-transmit classification scheme, offering limited flexibility in communication strategies. Furthermore, its hybridization of intermediate and late fusion is merely a simple superposition based on confidence scores, and it introduces additional bandwidth overhead.

III. Method

A. Problem Definition

Let $\mathcal{N} = \{1, ..., N\}$ be a set of collaborative agents. Each agent $i \in \mathcal{N}$ acquires a local observation $\mathcal{O}_i$ (e.g., LiDAR point cloud) and has access to its own learnable perception model parameterized by ϕ_i. In a collaborative setting, an ego-agent e aims to produce the most accurate detection results R_{final} by fusing its own information with collaborative information received from other agents.

CaCP introduces a novel context-adaptive communication scheme. Instead of transmitting monolithic data, each collaborator agent $i \neq e$ generates a structured, heterogeneous

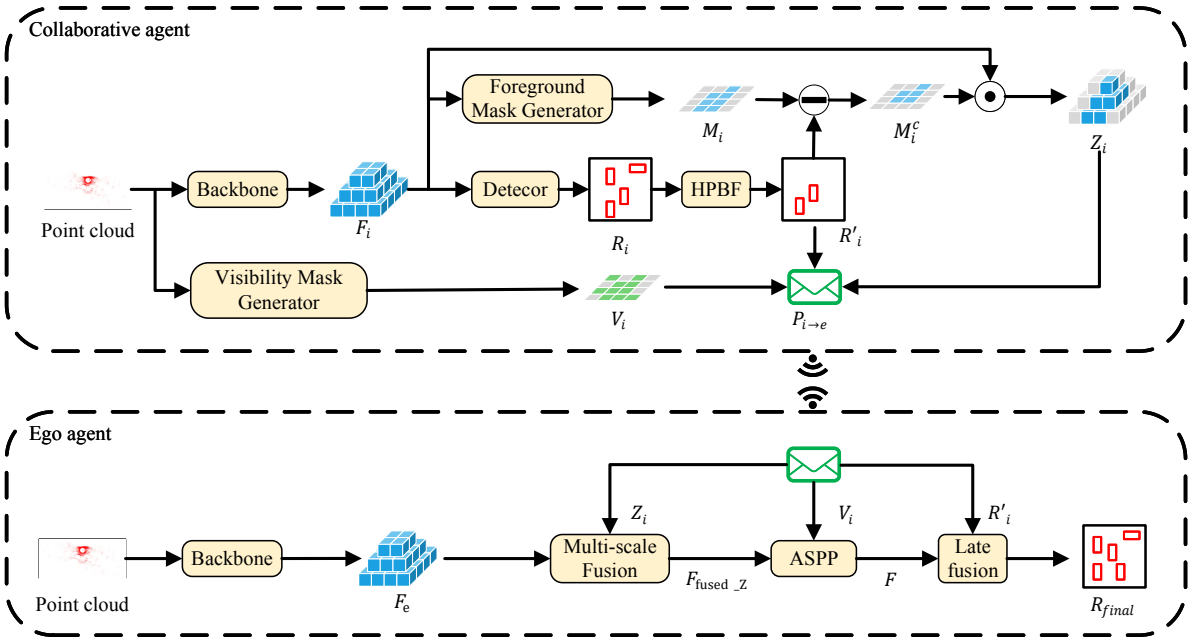

Fig. 2. The architecture of the proposed CaCP framework. CaCP enables collaborative agents to adaptively select information for transmission based on contextual information. The ego agent then employs a hybrid fusion strategy, performing feature-level fusion for intermediate collaboration regions and result-level fusion for late collaboration regions to generate the final detection.

message packet $P_{i\to e}$ based on a contextual partitioning of its own FoV. This message contains a tuple:

$$P_{i\to e} = \{Z_i, R'_i, V_i\}, \tag{1}$$

where Z_i represents sparse intermediate features from low-precision foreground regions, R'_i is a set of high-precision detection results from high-precision regions (for late fusion), and V_i is a compact visibility mask for the background. The generation of this message is governed by the CaCP framework, which relies on the agent's internal perception model.

The overall goal is to optimize the parameters $\Theta = \{\theta_e, \{\phi_i\}_{i\neq e}\}$ of the entire multi-agent system, where θ_e are the parameters of the ego-agent's fusion and detection modules, and ϕ_i are the parameters of the collaborator's perception and packaging modules. The optimization objective is to maximize the final detection accuracy $\mathscr{A}$, while the total communication cost $\sum_{i\neq e} \text{size}(P_{i\to e})$ is not larger than a certain communication budget, where $\text{size}(\cdot)$ measures the data volume in bits.

B. Overall Architecture

While prior works improve communication efficiency by selecting foreground regions, their uniform strategies remain suboptimal. They either expend bandwidth on well-perceived objects or discard valuable negative evidence from background areas. To address these limitations, CaCP operates as a hierarchical, context-adaptive framework. It assesses spatial content and local detection quality to dynamically select the most suitable communication strategy: feature-level intermediate fusion, detection-level late fusion, or minimalist visibility map sharing.

As depicted in Figure 2, the CaCP pipeline consists of two main stages: 1) Context-Aware Message Packing, and 2) Information Fusion and Detection. Before transmission, each collaborating agent i partitions its FoV into distinct contexts (low-precision foreground, high-precision foreground, and background) and packs a corresponding heterogeneous message $P_{i\to e}$ tailored to these contexts. After the ego-agent e receives these packets, it performs a hybrid fusion process that leverages both feature-level and object-level information to produce the final detection results.

C. Context-Aware Message Packing

This module intelligently analyzes the agent's local perception to construct the bandwidth-efficient message packet $P_{i\to e}$. The core idea is to adapt communication strategy based on the agent's perceptual precision: regions where the agent already achieves high-precision detection require only lightweight detection results sharing (late fusion), while uncertain or poorly localized regions necessitate transmitting rich intermediate features (intermediate fusion).

Our approach operates through a three-stage contextualization pipeline: (1) **Foreground-Background Partitioning** extracts potentially foreground regions using classification confidence; (2) **High-Precision Foreground Identification** further filters out well-detected objects suitable for late fusion using a dedicated quality assessment module; (3) **Background Characterization** provides compact visibility information to disambiguate empty and occluded regions. This hierarchical design enables CaCP to dynamically balance between communication efficiency and perception accuracy based on scene complexity and local detection quality.

Each collaborative agent first extracts multi-scale BEV features F_i from the point cloud via the backbone network and performs local object detection to obtain candidate results $R_i = \{r_1, r_2, ..., r_{N_i}\}$, where N_i is the number of detected objects. These serve as the foundation for subsequent context-aware message construction.

1) Foreground-Background Partitioning.: The first level of contextualization is to distinguish potentially object-rich foreground regions from empty background. Transmitting features for the entire BEV grid is usually highly redundant. Following Where2comm [11], we leverage the spatial confidence map produced by the classification head:

$$S_i = f_{\text{conf}}(F_i), \tag{2}$$

where f_{conf} is the classification head, S_i is the confidence map generated by this head. This map inherently reflects the likelihood of an object's presence. By applying a threshold τ_f, we generate a binary **Foreground Mask (FM)** M_i:

$$M_i(h,w) = \begin{cases} 1, & \text{if } S_i(h,w) > \tau_f, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where (h,w) denotes the spatial coordinates on the BEV plane. This mask M_i identifies all regions requiring further collaborative attention.

2) High-Precision Foreground Identification.: Simply transmitting all foreground features ($F_i \odot M_i$) may still be suboptimal. The reason is that for objects the local agent can already detect with high precision, sharing detailed intermediate features is unnecessary. Lightweight detection results (i.e., late fusion) suffice. However, a naive reliance on the detector's raw confidence score is misleading, as it often serves as a poor proxy for true localization quality. This discrepancy can lead to suboptimal fusion decisions, such as failing to request features for a poorly localized object that was assigned a high confidence score.

To address this issue, we propose a dedicated **High-Precision Bounding Box Filter (HPBF)** module designed to explicitly predict the localization quality of a detection. As illustrated in Figure 3, the HPBF is a lightweight network that takes a detected bounding box's parameters and its corresponding local BEV features as input. The BEV features are first refined through a channel attention module to emphasize the most informative channels for quality assessment. The attended features are then flattened and concatenated with encoded geometric parameters. The concatenated feature vector is subsequently normalized using layer normalization and fed into a MLP to predict the quality score $Q_a \in [0,1]$, which serves as a proxy for the bounding box's Intersection over Union (IoU) with the ground truth. This design allows the HPBF to learn a direct correlation between feature patterns, bounding box's geometry, and the final localization accuracy, making it far more reliable than the raw detection confidence.

Formally, for each bounding box $r_a \in R_i$, the HPBF computes its quality score Q_a as:

$$Q_a = f_{\text{HPBF}}(r_a, \text{crop}(F_i, r_a)), \tag{4}$$

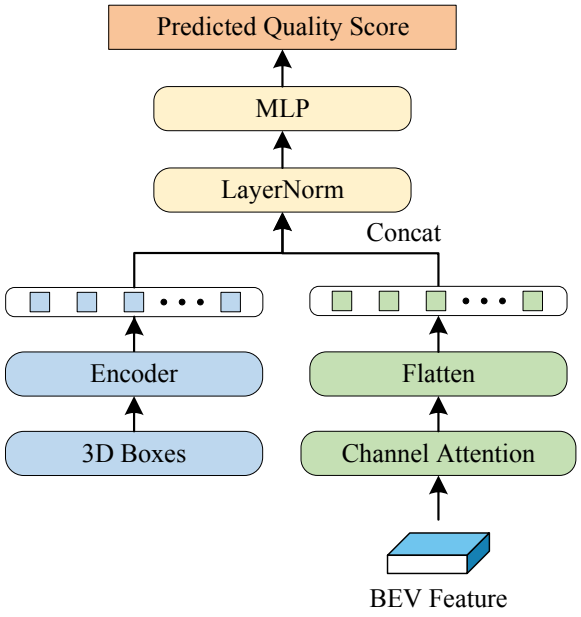

Fig. 3. Architecture of the HPBF module, which takes a detected bounding box's parameters and corresponding local BEV features as input to predict quality scores Q_a for bounding box filtering. The module applies channel attention to BEV features, concatenates them with encoded geometric parameters, and uses an MLP for quality prediction.

where $\text{crop}(F_i, r_a)$ extracts the feature patch corresponding to bounding box r_a. A bounding box is deemed high-precision for late fusion if $Q_a > \tau_q$, forming the set R_i'. The spatial locations of these high-precision bounding boxes are then subtracted from the initial foreground mask M_i to create the final collaboration mask M_i^c for intermediate fusion:

$$M_i^c = M_i \setminus \{ \bigcup_{r \in R_i'} \text{Mask}(r) \}, \tag{5}$$

where $\text{Mask}(r)$ is a binary mask of the area covered by bounding boxes r. The features selected for transmission are thus $Z_i = F_i \odot M_i^c$.

Training. The HPBF is trained in a decoupled fashion after the main perception network has converged and been frozen. To train the HPBF, each data batch is first passed through the frozen backbone to dynamically generate proposals and their corresponding features. These on-the-fly outputs, along with their calculated ground-truth IoUs, provide the supervision for the HPBF. It is trained as a binary classifier using focal loss, with proposals having IoU ¿ 0.7 as positive samples. This online training strategy avoids the need for a large, pre-generated dataset and allows the HPBF to learn from the exact feature distribution of the final perception model.

3) Background Characterization via Visibility Mask.: The masks M_i and M_i^c effectively eliminate collaboration on background regions. However, this creates ambiguity for the ego-agent: is a region empty because agent i saw nothing, or because it was occluded from agent i's view? This "negative evidence" is crucial for suppressing false positives. To resolve this ambiguity with minimal bandwidth cost, we

introduce a simple binary **Visibility Mask (VM)** V_i. It is generated by projecting agent i's raw point cloud $\mathcal{O}_i$ onto the BEV grid and marking cells with at least four points as "visible", an empirical threshold effective for filtering out sporadic noise points while retaining meaningful surface information.

$$V_i(h,w) = \begin{cases} 1, & \text{if } |\text{PointsInCell}(h,w)| > 3, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This visibility mask, when compressed, is extremely lightweight (a few thousand bits) yet provides valuable information about the observable space of each collaborator.

4) The Essence of Context-Adaptation.: The interplay of these modules embodies CaCP's context-adaptive nature. In a simple scene with clearly visible vehicles, the single-agent detector performs well. The HPBF identifies most bounding boxes as high-precision, i.e., $|R_i'|$ is large, causing the intermediate feature mask M_i^c to become very sparse. The collaboration automatically trends towards a highly efficient late fusion scheme. Conversely, in a complex and occluded scene, single-agent performance drops. The HPBF finds few high-precision bounding boxes, i.e., $|R_i'|$ is small, making M_i^c dense. The system automatically allocates more bandwidth to transmit rich intermediate features (Z_i), ensuring higher collaborative perception accuracy. This dynamic adjustment based on perceptual uncertainty is the key to CaCP's superior accuracy-bandwidth trade-off.

D. Information Fusion and Detection

The ego-agent e receives the heterogeneous packets $\{P_{i\to e}\}$ from its collaborators and performs a two-stage fusion process.

First, it conducts **Feature-Level Fusion**. The sparse intermediate features $\{Z_i\}$ from collaborators are fused with the ego-agent's own dense features F_e. Fusing only single-level BEV feature maps may lead to insufficient integration. CoAlign [23] addressed this issue by proposing a layer-by-layer fusion scheme for multi-scale features, which effectively resolves the inadequate fusion problem. We also adopt this approach for feature fusion in CaCP. Building on the demonstrated effectiveness of max fusion for BEV feature aggregation in prior work [12], [9], we adopt it as our primary channel-wise fusion operator:

$$F_{\text{fused\_Z}} = f_{\max}(F_e, \{T_{i\to e}(Z_i)\}_{i\neq e}), \tag{7}$$

where $T_{i\to e}$ is the coordinate transformation from agent i's frame to the ego-agent's frame and $f_{\max}$ is max fusion, selecting the maximum value of features from all collaborative agents.

Next, the collaborators' Visibility Masks $\{V_i\}$ are integrated. To effectively propagate these sparse binary cues into the dense feature map, we employ an Atrous Spatial Pyramid Pooling (ASPP) module [24]. The multi-scale receptive fields of ASPP allow the visibility information to influence a wider spatial context within the feature map without losing resolution, explicitly informing the network about regions confirmed to be empty by collaborators. The final fused feature map F is then:

$$F = \text{ASPP}(F_{\text{fused\_Z}}, \{T_{i\to e}(V_i)\}_{i\neq e}). \tag{8}$$

This map F is passed to the detection head to produce an initial set of collaborative detections, R_e.

Finally, **Result-Level Fusion** is performed. This stage combines the ego-agent's detections R_e with the high-precision bounding boxes $\{R_i'\}$ received directly from collaborators. A simple Non-Maximum Suppression (NMS) based on raw detection scores would be suboptimal. Instead, we perform NMS using the **HPBF-predicted quality scores** as the ranking criterion. This ensures that bounding boxes with higher predicted localization accuracy are prioritized, regardless of which agent they originated from. This quality-driven fusion process effectively leverages the strengths of both intermediate and late fusion, yielding the final, highly accurate detection set R_{final}.

IV. Experiments

A. Datasets and Experimental Settings

1) Datasets.: We conduct a comparative evaluation of CaCP against other approaches on the task of LiDAR-based 3D object detection, utilizing two distinct collaborative perception datasets: OPV2V[7] and V2V4Real [25]. These datasets encompass both simulated and real-world scenes. Specifically, OPV2V is a large-scale simulated benchmark containing 10,914 annotated frames, and V2V4Real is a real-world dataset collected over 410 km of driving, comprising 20,000 LiDAR frames sampled at 10Hz. The inclusion of both datasets allows for a comprehensive evaluation of CaCP's performance and generalization capabilities across simulated and real-world conditions.

2) Evaluation Metrics.: We evaluate the 3D detection performance using Average Precision (AP) at IoU thresholds of 0.5 and 0.7 (denoted as AP@0.5 and AP@0.7). All AP scores are reported in percentages. For communication efficiency, we measure the Communication Bandwidth (Mbps). This is calculated by first summing the data volume transmitted among all agents for all frames in the test set, and then dividing by the total duration of the test set sequence (assuming a 10Hz frame rate). For a collaborative agent i, the single-frame communication cost (in bits) for different data types is defined as follows:

Feature Representation: The volume for transmitting a feature map selected by an intermediate feature mask M_i^c is:

$$\text{Volume}_{\text{feat}} = H \times W \times C \times K_i \times 32, \tag{9}$$

where H, W, C are the height, width, and channel dimensions of the feature map, and we assume 32-bit floating-point precision, K_i is the proportion of selected elements in M_i^c.

Visibility Mask: The volume for the visibility mask, which is compressed by a factor of 16 before transmission, is:

$$\text{Volume}_{\text{mask}} = (H \times W \times 1)/16, \tag{10}$$

where each element is a single bit before compression.

TABLE I

COMPARISON OF PERCEPTION ACCURACY AND COMMUNICATION VOLUME OF DIFFERENT METHODS ON OPV2V AND V2V4REAL. METHODS MARKED WITH * USE A FOREGROUND MASK. TO ENSURE A FAIR COMPARISON ON COMMUNICATION OVERHEAD, WE DYNAMICALLY ADJUST THE FOREGROUND SELECTION THRESHOLD FOR EACH METHOD SO THAT THE PROPORTION OF SELECTED FOREGROUND AREA MATCHES THAT OF THE MOST COMMUNICATION-EFFICIENT BASELINE (E.G., MATCHING COSDH*'S AREA RATIO).

Dataset	OPV2V			V2V4Real		
Method	AP@0.5	AP@0.7	BD (Mb/s)	AP@0.5	AP@0.7	BD (Mb/s)
Late Fusion	89.06	86.24	0.29	57.02	32.60	0.12
Intermediate Fusion	93.81	87.99	1865.36	39.17	14.44	1999.61
V2X-ViT	94.33	89.21	1790.75	68.29	40.85	1333.08
Where2comm*	95.03	89.83	44.73	65.09	39.88	99.25
CoAlign	94.79	89.52	1865.36	66.79	37.70	1388.62
CoSDH*	95.75	90.25	42.27	68.13	40.99	**94.58**
CaCP*	**96.27**	**93.03**	**34.90**	**69.79**	**41.15**	99.19

Detection Results (bounding boxes): The volume for transmitting N detected bounding boxes is:

$$\text{Volume}_{\text{det}} = N \times (7+1) \times 32, \tag{11}$$

where each bounding box is represented by a 7-dimensional vector (containing position, shape, and yaw angle information) and a 1-dimensional quality score, both with 32-bit precision.

3) Implementation.: Our experiments are based on the OpenCOOD [7] framework. We use a PointPillar [26] encoder with a grid size of (0.4m, 0.4m, 4m) and a maximum of 32 points per pillar. For the foreground mask threshold τ_f, we adopt the standard setting of 0.01 following Where2comm and CoSDH. However, in benchmark comparisons, we replace the fixed threshold with a proportional selection mechanism (selecting a fixed percentage of foreground areas) to ensure fair comparison across methods with different foreground detection sensitivities. The threshold τ_q in HPBF is set to 0.8 (OPV2V) and 0.6 (V2V4Real), determined through a grid search on the validation set. We provide sensitivity analysis of this parameter in the Ablation Studies. The experiments employ the Adam [27] optimizer with a cosine annealing learning rate schedule with warmup, and models are trained for 40 epochs to ensure convergence. The maximum number of collaborating agents is set to 5. All other parameters are consistent with the OpenCOOD framework. All experiments were conducted on eight NVIDIA GeForce RTX 3090 GPUs. For a fair comparison, no feature compression techniques were applied to any of the methods, and all models were trained from scratch under the same configuration and environment.

4) Baselines.: We compare our proposed CaCP framework against representative baselines from three complementary categories:

Classical mid- and late-fusion baselines. We adopt the standard intermediate fusion and late fusion implementations provided by the OpenCOOD benchmark [7]. These methods serve as widely recognized references for cooperative perception and allow us to evaluate the fundamental effectiveness of CaCP over traditional fusion paradigms.

Accuracy-oriented intermediate fusion methods (without communication constraints). We include V2X-ViT [10] and CoAlign [23], both of which are state-of-the-art collaborative perception frameworks focusing on improving detection accuracy and robustness, without explicitly optimizing communication bandwidth.

Communication-efficient frameworks. We further compare against Where2Comm [11] and CoSDH [12], two recent approaches that reduce bandwidth usage via feature selection. Notably, CoSDH represents the most recent state-of-the-art in this category.

Remarks. We exclude certain feature-compression-based approaches (e.g., CodeFilling) from direct comparison, as they are orthogonal to CaCP. In principle, these methods can be integrated with CaCP's feature selection module to further compress the selected features, which is outside the scope of our current evaluation.

B. Quantitative Evaluation

1) Benchmark Comparison.: Table I presents a comparison of the collaborative perception accuracy and communication volume of CaCP against previous approaches on different datasets. The results demonstrate that, without using any additional compression methods, CaCP achieves the highest perception accuracy on the OPV2V dataset while using only 83% of the bandwidth required by the second most communication-efficient method. Compared to methods that are not communication-efficient, the bandwidth consumption of CaCP is merely around 2%. On the V2V4Real dataset, CaCP also attains the highest perception accuracy, although it does not achieve the lowest communication overhead. This is because the higher levels of noise and error present in the real-world dataset degrade the model's intrinsic detection performance, resulting in lower average precision for bounding boxes generated by single agents. Consequently, CaCP adaptively reduces the region designated for late collaboration, which in turn increases the communication overhead. This also highlights the flexibility of our method: it does not indiscriminately reduce communication overhead. Instead, it transmits more valuable information to ensure perception accuracy when detection quality is compromised.

C. Qualitative Evaluation

Figure 4 shows a qualitative comparison between CaCP and other methods on the OPV2V and V2V4Real datasets. Compared to Where2comm, CaCP achieves a higher recall.

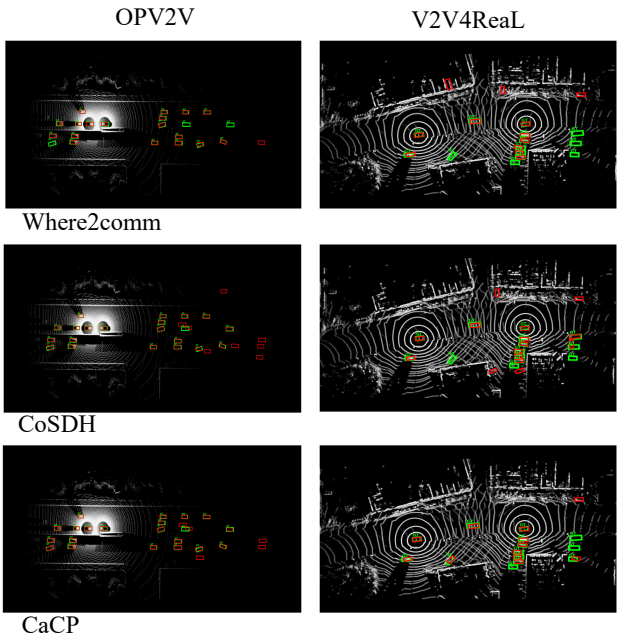

Fig. 4. Visualization of detection results on the OPV2V and V2V4Real datasets. Green represents ground-truth bounding boxes, and red represents detected bounding boxes.

In comparison with CoSDH, CaCP demonstrates a comparable recall but attains higher precision and yields fewer false positives on the OPV2V dataset.

TABLE II

ABLATION STUDY OF THE MODULES IN CACP ON THE OPV2V DATASET. HERE, FM DENOTES THE USE OF A FOREGROUND MASK FOR FOREGROUND PARTITIONING, VM REPRESENTS THE VISIBILITY MASK FOR BACKGROUND REGIONS, AND HPBF SIGNIFIES THE HIGH-PRECISION FOREGROUND IDENTIFICATION USING THE HPBF MODULE. ADDITIONALLY, CONF REPRESENTS USING CLASSIFICATION CONFIDENCE INSTEAD OF HPBF. THE FIRST ROW REPRESENTS THE FULL INTERMEDIATE FUSION BASELINE.

FM	VM	HPBF	Conf.	AP@0.5	AP@0.7	BD(Mb/s)
				95.81	92.45	1865.36
✓				95.52	92.23	44.89
✓	✓			95.75	92.53	44.94
✓	✓	✓		96.27	93.03	34.90
✓	✓		✓	95.73	91.72	37.64

D. Ablation Studies

Table II presents the results of an ablation study on the components of CaCP using the OPV2V dataset. The results indicate that collaborating only on foreground regions can significantly reduce communication overhead, albeit with a slight loss in accuracy. The introduction of the visibility mask for background regions compensates for this accuracy loss at an extremely low communication cost (1/1000 of the full feature transmission). The hybrid intermediate-late collaboration enabled by the HPBF module not only reduces bandwidth consumption by approximately 23% but also maintains, and even slightly improves, perception accuracy. This improvement is primarily attributed to the high-quality late collaboration component of the hybrid scheme. Lastly, while simply using classification confidence scores in place of the quality score Q_a output by the HPBF module still reduces some communication overhead, it incurs a significant loss in accuracy, which unequivocally demonstrates the necessity of the HPBF module. In summary, these results systematically validate our FoV decomposition strategy. The effectiveness of the visibility mask and the HPBF module confirms that partitioning the FoV into four distinct categories allows for a more granular and efficient collaboration approach than a coarse foreground-background division.

TABLE III

SENSITIVITY ANALYSIS OF THE QUALITY THRESHOLD τ_q IN OUR HPBF MODULE. WE EVALUATE THE IMPACT OF DIFFERENT τ_q VALUES ON THE AP@0.5 AND AP@0.7 USING THE OPV2V VALIDATION DATASET. BOLD INDICATES THE CHOSEN VALUE FOR OUR FINAL CONFIGURATION.

τ_q	0.6	0.7	0.75	**0.8**	0.9	1.0
AP@0.5	95.75	95.78	95.80	95.83	95.76	95.20
AP@0.7	91.60	91.72	91.78	91.82	91.98	91.08
BD(Mb/s)	67.81	71.84	74.30	77.43	91.69	102.38

Sensitivity Analysis of τ_q. The threshold τ_q in our proposed HPBF governs the critical trade-off between detection accuracy and communication bandwidth by dynamically allocating regions to either late or intermediate fusion. We tune this hyperparameter on the validation sets. As shown in Table III for OPV2V, increasing τ_q initially improves AP@0.5 as it filters for higher-quality detections for late fusion. However, this simultaneously forces more regions into the bandwidth-intensive intermediate fusion pathway. For $\tau_q > 0.8$, performance declines as the pool of high-confidence detections becomes too sparse to be effective. Consequently, we identify $\tau_q = 0.8$ as the optimal balance point for OPV2V. While $\tau_q = 0.9$ achieves a marginal gain in AP@0.7 (+0.16), it incurs a significant 18% increase in bandwidth. Following an identical procedure for the V2V4Real dataset, we determine its optimal threshold to be $\tau_q = 0.6$. This lower threshold is consistent with the generally lower quality of detections in its challenging real-world scenes compared to the simulated OPV2V. These tuned values are used in all other experiments.

V. CONCLUSION

This paper proposes CaCP, a context-adaptive communication method for collaborative perception. Based on the varying contexts of different perceptual regions, CaCP selectively transmits different types of collaborative information and applies distinct collaboration strategies, thereby avoiding communication redundancy and maximizing bandwidth utilization. In particular, our policy of transmitting only detection results for objects that a single agent can perceive with high precision naturally establishes a spatially hybrid intermediate-late collaboration scheme. Experiments across multiple datasets demonstrate that CaCP achieves a superior trade-off between accuracy and bandwidth. It can also flexibly adjust communication overhead in response to

the current perception performance, thereby ensuring that perception accuracy is maximized.

REFERENCES

[1] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[2] X. Gao, X. Zhang, Y. Lu, Y. Huang, L. Yang, Y. Xiong, and P. Liu, "A survey of collaborative perception in intelligent vehicles at intersections," *IEEE Transactions on Intelligent Vehicles*, pp. 1–20, 2024.

[3] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 1852–1864, 2022.

[4] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, 2019, pp. 514–524.

[5] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, 2022, pp. 21 329–21 338.

[6] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[7] 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)*. Philadelphia, PA, USA: IEEE Press, 2022, pp. 2583–2589.

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

[9] 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*, vol. 12347. Cham: Springer International Publishing, 2020, pp. 605–621.

[10] 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 *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*. Berlin, Heidelberg: Springer-Verlag, 2022, pp. 107–124.

[11] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022, pp. 4874–4886.

[12] J. Xu, Y. Zhang, Z. Cai, and D. Huang, "CoSDH: Communication-efficient collaborative perception via supply-demand awareness and intermediate-late hybridization," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 6834–6843.

[13] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proceedings of the 5th Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[14] Y. Liu, T. Wang, X. Zhang, and J. Sun, "PETR: Position embedding transformation for multi-view 3D object detection," in *Computer Vision – ECCV 2022*, vol. 13687. Cham: Springer Nature Switzerland, 2022, pp. 531–548.

[15] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, 2017, pp. 77–85.

[16] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2020, pp. 10 526–10 535.

[17] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1080–1089.

[18] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2774–2781.

[19] 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*. Arlington Virginia: ACM, 2019, pp. 88–100.

[20] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Proceedings of The 6th Conference on Robot Learning*. PMLR, 2023, pp. 989–1000.

[21] 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.

[22] Y. Hu, J. Peng, S. Liu, J. Ge, S. Liu, and S. Chen, "Communication-efficient collaborative perception via information filling with codebook," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, 2024, pp. 15 481–15 490.

[23] 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)*, 2023, pp. 4812–4818.

[24] J. Li, C. Luo, and X. Yang, "PillarNeXt: Rethinking network designs for 3D object detection in LiDAR point clouds," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 17 567–17 576.

[25] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, H. Yu, B. Zhou, and J. Ma, "V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 712–13 722.

[26] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 689–12 697.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.