What2comm: Towards Communication-efficient Collaborative Perception via Feature Decoupling

Kun Yang*
Academy for Engineering and Technology, Fudan University
kunyang20@fudan.edu.cn

Dingkang Yang*
Academy for Engineering and Technology, Fudan University
dkyang20@fudan.edu.cn

Jingyu Zhang
Academy for Engineering and Technology, Fudan University

Hanqi Wang
Academy for Engineering and Technology, Fudan University

Peng Sun
Duke Kunshan University

Liang Song†
Academy for Engineering and Technology, Fudan University

ABSTRACT

Multi-agent collaborative perception has received increasing attention recently as an emerging application in driving scenarios. Despite advancements in previous approaches, challenges remain due to redundant communication patterns and vulnerable collaboration processes. To address these issues, we propose *What2comm*, an end-to-end collaborative perception framework to achieve a trade-off between perception performance and communication bandwidth. Our novelties lie in three aspects. First, we design an efficient communication mechanism based on feature decoupling to transmit exclusive and common feature maps among heterogeneous agents to provide perceptually holistic messages. Secondly, a spatio-temporal collaboration module is introduced to integrate complementary information from collaborators and temporal ego cues, leading to a robust collaboration procedure against transmission delay and localization errors. Ultimately, we propose a common-aware fusion strategy to refine final representations with informative common features. Comprehensive experiments in real-world and simulated scenarios demonstrate the effectiveness of What2comm.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; *Neural networks*; • **Information systems** → *Multimedia streaming*.

KEYWORDS

3D object detection, collaborative perception, feature decoupling

ACM Reference Format:
Kun Yang, Dingkang Yang, Jingyu Zhang, Hanqi Wang, Peng Sun, and Liang Song. 2023. What2comm: Towards Communication-efficient Collaborative Perception via Feature Decoupling. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611699

*indicates equal technological and writing contributions in no particular order. The work was done when Dingkang Yang was at the Institute of Meta-Medical, IPASS.
†indicates the corresponding author.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611699

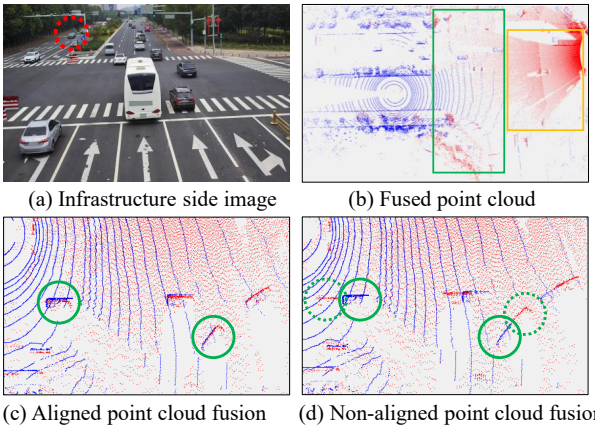


Figure 1: (a) and (b) show the image and fused point cloud of the scene containing an ego vehicle (red circle) and infrastructure, respectively. Green and orange boxes denote common and exclusive perception regions, respectively. The comparison of (c) and (d) shows the point cloud fusion errors due to transmission delay.

1 INTRODUCTION

Precise environmental perception is essential for ensuring the driving safety of Autonomous Vehicles (AVs). Benefiting from advances in deep learning-based technologies [3, 5, 12, 14, 18–21, 29, 30, 34, 35, 39, 42, 44–46, 48, 59], numerous studies are devoted to optimizing the accuracy of in-vehicle vision applications, including object detection [16, 49, 50, 58] and instance segmentation [10, 26, 27]. However, this kind of single-agent perception paradigm is inevitably restricted by several natural conditions, such as occlusion [54], limited detection range [55], and severe weather [57], making it more challenging to achieve robust vehicular perception. Recently, multi-agent collaborative perception [15, 22, 32, 38, 40, 41] has developed as a promising solution to overcome the above physical limitations. This novel perception system facilitates information sharing among on-road agents via Vehicle-to-Everything (V2X)

communication, leading to a more holistic perception of surrounding driving scenarios. Based on the emerging collaborative perception datasets [40, 41, 52], existing efforts seek a trade-off between performance and bandwidth via seminal communication and collaboration mechanisms. Despite recent advancements, challenges remain due to various collaboration noises, including transmission delay [13, 53], localization errors [28], and agent heterogeneity [40].

As for communication mechanisms, current feature compression-based methods [15, 32, 40] ignore the spatial heterogeneity of feature maps. In addition, spatial filtering-based algorithms [6] rely on the trained confidence maps, which may only focus on the high-confidence areas and fail to extract the complementary information among agents. Meanwhile, it is difficult for existing strategies to handle the data discrepancies caused by the agent heterogeneity in the sensor type and installation height [37, 40]. Figures 1(a)&(b) demonstrate the collaborative perception scenario involving two agents and the fused point cloud. Intuitively, the orange box denotes the exclusive perception region of the collaborator, which can serve as complementary information for the ego vehicle to expand the detection range and complete the occluded areas. The overlapping perception range of the green box maintains the common semantic information and is beneficial to bridge the data distribution gap [2]. Overall, considering exclusive and common information can facilitate efficient and pragmatic communication patterns.

Moreover, temporal asynchrony caused by transmission delay potentially degrades the collaboration performance [13, 53]. Figures 1(c)&(d) show the fused point clouds in the time-synchronous and time-asynchronous cases, respectively. The moving vehicles inside the green circles produce the fusion errors of the two-side point clouds due to the time delay, resulting in position misalignment and false detection results. However, the existing single-frame perception pattern restricts delay compensation methods and leads to performance bottlenecks. Also, the localization errors of agents may cause feature misalignment and harm the detection precision of per-agent/location message fusion efforts [15, 22, 31, 32, 40].

Motivated by the above observations, we propose *What2comm*, a unified communication-efficient multi-agent collaborative perception framework to address the existing challenges in an end-to-end manner. From Figure 2, What2comm contains three core components: i) a Decoupling-based Communication Mechanism (DCM), which captures the exclusive and common representations among distinct agents via feature disentanglement to determine *what* messages to communicate. DCM provides a communication-efficient information sharing pattern through feature specificity and consistency supervision; ii) a Spatio-Temporal Collaboration Module (STCM), which aggregates perceptually complementary information from exclusive feature maps shared by collaborators and ego-centered temporal semantics. STCM mitigates feature misalignment due to transmission delay and localization errors by joint spatio-temporal modeling; iii) a Common-Aware Fusion (CAF) strategy, which extracts high-dimensional information from collaborator-shared common representations to eliminate the data distribution gap across agents. Benefiting from the above-customized communication and collaboration components, What2comm takes a solid step progressed to a communication-efficient and noise-robust collaborative perception system. We evaluate the performance of What2comm on several public LiDAR-based collaborative 3D object detection datasets, including DAIR-V2X [52], V2XSet [40], and OPV2V [41]. Comprehensive experiment results show that the proposed What2comm achieves better performance-bandwidth trade-off than the state-of-the-art (SOTA) collaborative perception works. The main contributions are summarized as follows:

- We propose What2comm, a communication-efficient multi-agent collaborative perception framework. Our framework outperforms previous approaches on real-world and simulated datasets by addressing various collaboration interferences, including communication noises, transmission delay, and localization errors, in an end-to-end manner.
- We present a novel decoupling-based communication mechanism to promote comprehensive and pragmatic information transmission among heterogeneous agents.
- We design a spatio-temporal collaboration module to effectively integrate collaborators' exclusive representations and historical ego cues. Meanwhile, a common-aware fusion strategy is introduced to fuse the collaborators' common features and reinforce the final representations.

2 RELATED WORK

2.1 Multi-agent Collaborative Perception

Multi-agent collaborative perception [6, 15, 22, 23, 31, 32, 38, 40, 51] as an emerging application for LiDAR-based 3D object detection is only in its infancy. There are several well-designed datasets to promote the development of collaborative perception areas, such as DAIR-V2X [52], OPV2V [41], and V2XSet [40]. In this case, various seminal methods are proposed to achieve a trade-off between perception performance and communication bandwidth. For instance, V2VNet [32] utilized a spatially aware graph neural network to aggregate the information received from all the nearby self-driving vehicles. After that, DiscoNet [15] introduced a teacher-student framework to aggregate the benefits of early collaboration with holistic views and intermediate collaboration with a single view. V2X-ViT [40] presented a unified transformer architecture to capture the heterogeneous nature of V2X systems with strong robustness against various noises. More recently, Where2comm [6] employed a confidence-aware pattern to guide agents to focus on sharing spatially critical information. Unfortunately, most methods follow only a single-frame perception paradigm which suffers from the data sparsity dilemma of 3D point clouds. In this paper, we propose a *spatio-temporal collaboration module* to progressively learn the spatial semantics and temporal dynamics among agents.

2.2 Feature Decoupled Learning

Decoupled learning aims to explore diverse attributes in heterogeneous high-dimensional spaces to capture distinct features [43]. Early efforts at feature decoupling are based on auto-encoders [1] or generative adversarial networks [24]. Subsequently, FactorVAE [8] proposed to decouple by encouraging the representation to be factorial and independent across the dimensions. DVR [33] introduced a variational lower bound to estimate the posterior and optimize the latent variable space, aiming at disentangling the infrared and visible face representations. Moreover, MFSA [47] presented a double-discriminator decoupling strategy to supervise the learning

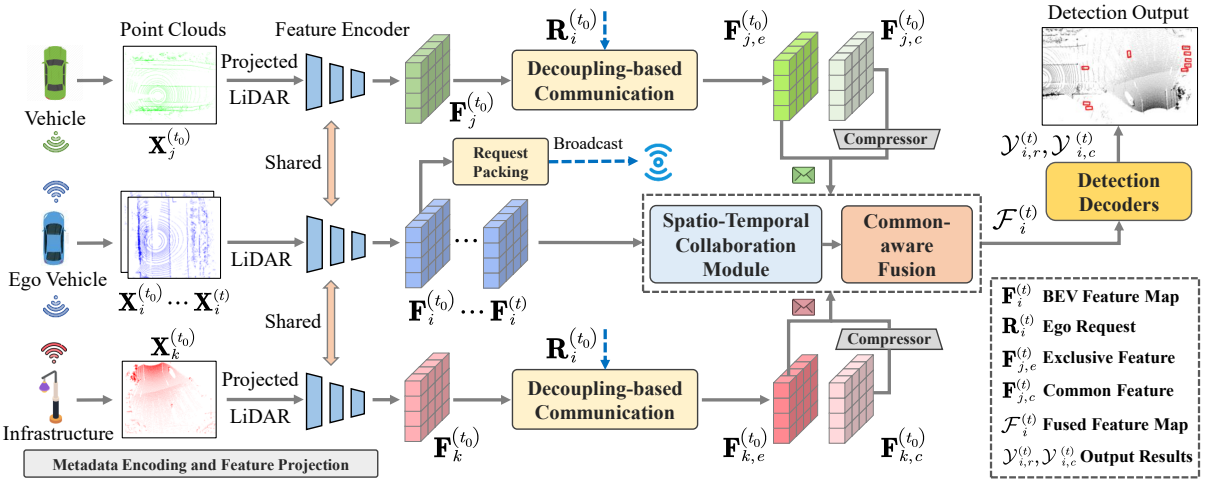


Figure 2: Overall architecture of the proposed What2comm. This framework comprises five parts: metadata encoding and feature projection, decoupling-based communication mechanism, spatio-temporal collaboration module, common-aware fusion strategy, and detection decoders. Each part is detailed in Section 3. The time delay τ (*i.e.*, $\tau = t - t_0$) is considered.

of different representations. In comparison, this paper designs a *decoupling-based communication mechanism* to achieve pragmatic information transmission in multi-agent collaborative perception.

3 APPROACH

This paper aims to develop a communication-efficient multi-agent collaborative perception system to improve the ego agent's perception ability. Formally, let $\boldsymbol{X}_i^{(t)}$ and $\tilde{\boldsymbol{\mathcal{Y}}}_i^{(t)}$ in a scene with N agents represent the local point cloud observation and perception supervision from i-th agent at timestamp t, respectively. The objective of the overall collaborative system is to maximize the LiDAR-based 3D detection performance under a total communication budget B:

$$\arg\max_{\theta, \boldsymbol{F}_{j\to i}} \sum_i \varrho(\Psi_\theta(\boldsymbol{X}_i^{(t)}, \{\boldsymbol{F}_{j\to i}^{(t_0)}\}_{j=1}^N), \tilde{\boldsymbol{\mathcal{Y}}}_i^{(t)}), \text{ s.t.} \sum_j |\boldsymbol{F}_{j\to i}^{(t_0)}| \le B, \tag{1}$$

where $\varrho(\cdot,\cdot)$ stands for the perception evaluation metric, Ψ_θ is a collaborative system parameterized by θ, and $\boldsymbol{F}_{j\to i}^{(t_0)}$ is the collaboration message transmitted from the j-th agent to the i-th agent at time delay τ-aware moment t_0 *w.r.t.* $t_0 = t - \tau$. Figure 2 illustrates the overall architecture of the proposed system framework. The remainder of Section 3 details the following vital components.

3.1 Metadata Encoding and Feature Projection

In the initial stage of collaboration, the necessary metadata, such as poses and extrinsics, are shared among each agent $j \in \{1, ..., N\}$. An agent is identified as the ego agent (i) while the other connected agents (*e.g.*, infrastructures and AVs) act as collaborators in the communication connection. After receiving the ego agent's poses, the connected collaborators encode and project their local LiDAR point clouds to the ego agent's coordinate frame for better cross-agent collaboration. Additionally, we synchronize the ego agent's previous point cloud frames to the current coordinate. Then a feature encoder is used to convert the j-th agent's 3D point cloud into the BEV feature map as $\boldsymbol{F}_j^{(t)} = f_{enc}(\boldsymbol{X}_j^{(t)}) \in \mathbb{R}^{H\times W\times C}$, where $f_{enc}(\cdot)$ is the PointPillar [11] encoder shared among all agents, and H, W, C represents height, width, and channel, respectively.

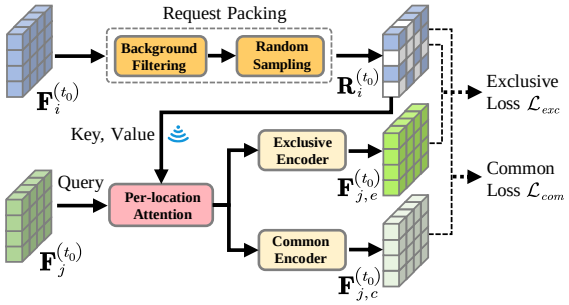


Figure 3: Overall structure of the proposed decoupling-based communication mechanism.

3.2 Decoupling-based Communication

In order to achieve efficient communication among multiple agents, most previous approaches utilized auto-encoders [15, 23] or spatial confidence maps [6, 7] to reduce the required transmission bandwidth. However, these communication strategies ignore the heterogeneity of the transmitted information due to configuration differences among agents [40], leading to sub-optimal solutions that subsequently affect collaboration performance. To this end, we present a Decoupling-based Communication Mechanism (DCM) to achieve an effective performance-bandwidth trade-off. The design philosophy of DCM is to learn the specificity and consistency of the transmitted feature representations. The details are as follows.

Request Packing. As Figure 3 shows, we first derive a request $\boldsymbol{R}_i^{(t_0)}$ based on the ego feature $\boldsymbol{F}_i^{(t_0)}$ to assist collaborators in feature decoupling. To make the ego request more compact and informative, we adopt the importance map to filter the background areas in

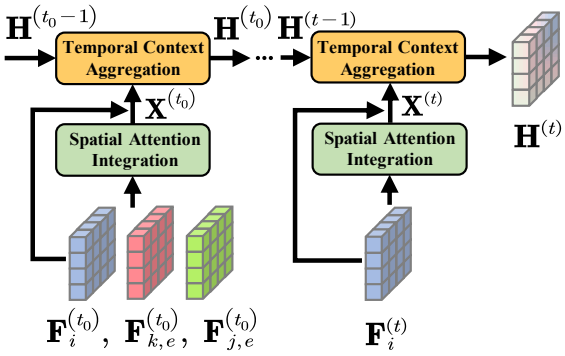


Figure 4: Overall structure of the proposed spatio-temporal collaboration module, including two core components: spatial attention integration and temporal context aggregation.

the ego feature. The importance generator $f_{gen}(\cdot)$ generates the importance map of $\boldsymbol{F}_i^{(t_0)}$ as follows:

$$S_i^{(t_0)} = \sigma \cdot \Phi_m(f_{gen}(\boldsymbol{F}_i^{(t_0)})) \in [0,1]^{H\times W}, \tag{2}$$

where σ is the sigmoid activation, and Φ_m is the channel-wise max pooling operation. Then, we uniformly sample N_r pixels from the filtered feature to form the ego request. Upon receiving the broadcasted request $\boldsymbol{R}_i^{(t_0)}$, the collaborators refine the local feature via per-location cross-attention [6] and obtain the exclusive and common features $\{\boldsymbol{F}_{j,e}^{(t_0)}, \boldsymbol{F}_{j,c}^{(t_0)}\}$ through two encoders. The following two constraints are proposed for feature decoupling supervision.

Specificity Constraint. Specificity constraint is used to supervise the extraction of the collaborators' exclusive representations, which can serve as complementary information for the ego agent to promote perspective complementation. In detail, we find the corresponding pixels of request $\boldsymbol{R}_i^{(t_0)}$ in $\boldsymbol{F}_{j,e}^{(t_0)}$ to form N_r pairs of features $\{z_{i/j}^n\}_{n=1}^{N_r}$, where $z_{i/j}^n \in \mathbb{R}^{1\times C}$. Then the $L2$-normalization is employed across the channel dimension. Naturally, the corresponding spatial regions of the ego request and exclusive features should contain diverse semantics. Therefore, we utilize the orthogonal distance to measure the distribution variance of a pair of features. The exclusive loss function is formulated as follows:

$$\mathcal{L}_{exc} = \frac{1}{(N-1)*N_r} \sum_{j\in\{1,\dots,N\}, j\neq i} \sum_{n=1}^{N_r} \|z_i^n \cdot (z_j^n)^T\|_F^2, \tag{3}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm.

Consistency Constraint. Minimizing the gap between $\boldsymbol{R}_i^{(t_0)}$ and $\boldsymbol{F}_{j,c}^{(t_0)}$ is beneficial in overcoming data discrepancies between collaborators and the ego agent and maintaining valuable and salient information in common locations. For this purpose, we employ the Central Moment Discrepancy (CMD) [56] distance metric to measure the distribution between two representations by matching their order-wise moment differences. Let $\hat{\boldsymbol{R}}_i^{(t_0)}$ and $\hat{\boldsymbol{F}}_{j,c}^{(t_0)}$ be bounded samples with respective probability distributions p and q on the interval $[a,b]^{\mathcal{B}}$. The central moment discrepancy

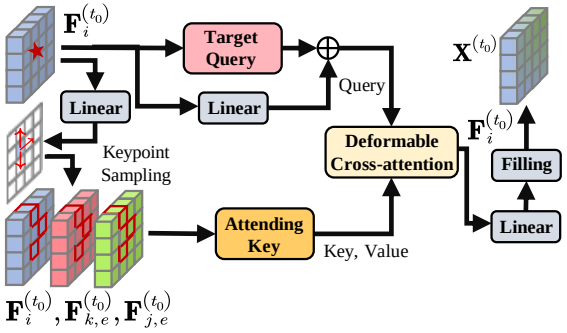


Figure 5: Overall structure of the proposed deformable cross-attention based spatial attention integration component.

regularizer CMD_M is expressed as an empirical estimation:

$$\begin{aligned} CMD_M(\boldsymbol{R}_i^{(t_0)}, \boldsymbol{F}_{j,c}^{(t_0)}) &= \frac{1}{|b-a|}\|\mathbb{E}(\hat{\boldsymbol{R}}_i^{(t_0)}) - \mathbb{E}(\hat{\boldsymbol{F}}_{j,c}^{(t_0)})\|_2 \\ &+ \sum_{m=2}^{M} \frac{1}{|b-a|^m}\|C_m(\hat{\boldsymbol{R}}_i^{(t_0)}) - C_m(\hat{\boldsymbol{F}}_{j,c}^{(t_0)})\|_2, \end{aligned} \tag{4}$$

where $\mathbb{E}(\hat{\boldsymbol{R}}_i^{(t_0)}) = \frac{1}{|\hat{\boldsymbol{R}}_i^{(t_0)}|}\sum_{\hat{\boldsymbol{r}}_i^{(t_0)}\in\hat{\boldsymbol{R}}_i^{(t_0)}} \hat{\boldsymbol{r}}_i^{(t_0)}$ is the empirical expectation vector of the feature $\hat{\boldsymbol{R}}_i^{(t_0)}$, $C_m(\hat{\boldsymbol{R}}_i^{(t_0)}) = \mathbb{E}\left((\hat{\boldsymbol{r}}_i^{(t_0)} - \mathbb{E}(\hat{\boldsymbol{R}}_i^{(t_0)}))^m\right)$ is the vector of all m^{th} order feature central moments of the coordinates of $\hat{\boldsymbol{R}}_i^{(t_0)}$, and $\hat{\boldsymbol{R}}_i^{(t_0)}/\hat{\boldsymbol{F}}_{j,c}^{(t_0)} = \aleph(\boldsymbol{R}_i^{(t_0)}/\boldsymbol{F}_{j,c}^{(t_0)})$. $\aleph(\cdot)$ refers to the adaptive max pooling operation. Here, we calculate each pair of the CMD metric between $\boldsymbol{R}_i^{(t_0)}$ from the ego agent and $\boldsymbol{F}_{j,c}^{(t_0)}$ from the j-th connected collaborator as the common loss:

$$\mathcal{L}_{com} = \frac{1}{N-1} \sum_{j\in\{1,\dots,N\}, j\neq i} CMD_M(\boldsymbol{R}_i^{(t_0)}, \boldsymbol{F}_{j,c}^{(t_0)}). \tag{5}$$

For the decoupled representations $\{\boldsymbol{F}_{j,e}^{(t_0)}, \boldsymbol{F}_{j,c}^{(t_0)}\}$, we obtain their corresponding importance maps using Equation (2) and filter the background to get sparse yet critical features. As Figure 2 shows, a 1×1 convolutional compressor compresses $\boldsymbol{F}_{j,c}^{(t_0)}$ along the channel dimension to reduce bandwidth. Then, the sparse exclusive features and compressed common features are transmitted to the ego agent.

3.3 Spatio-Temporal Collaboration Module

The collaboration module targets to enhance the ego agent's visual representation by aggregating complementary semantics from the exclusive features of collaborators and local historical observations. To this end, we implement a Spatio-Temporal Collaboration Module (STCM) to fuse delayed exclusive features $\boldsymbol{F}_{j,e}^{(t_0)} \in \mathbb{R}^{H\times W\times C}, j \in \{1,\dots,N\}$ and historical frames $\{\boldsymbol{F}_i^{(t_0)},\dots,\boldsymbol{F}_i^{(t)}\}$ in Figure 4. For the spatial fusion on each timestamp, we design a Spatial Attention Integration (SAI) based on deformable cross-attention [60] to attend each potential target with its rich spatial semantics across agents. Also, we propose a GRU-like component named Temporal Context Aggregation (TCA) to refine the meaningful object characteristics and flexibly incorporate the temporal context through a gate mechanism. When the collaborator-shared features are not synchronized,

STCM separately aligns the delayed collaborator features with the ego historical observations and passes aligned features into the SAI.

Spatial Attention Integration. The SAI component extracts spatial information from exclusive representations of collaborators. For input features, we obtain their corresponding importance maps using Equation (2) and select local maximum elements as queries since they potentially contain the targets. Subsequently, we collect the target queries from all agents to form the query embedding, which can actively guide the SAI sub-module to focus on foreground objects. As Figure 5 shows, a linear layer learns 2D spatial offsets $\{\Delta p_v \mid v \in 1, ..., N_v\}$ for each query p and samples the keypoints at $p + \Delta p_v$. The features of these keypoints are extracted as the attending features. After adding the position embedding learned by a linear layer, we obtain the enhanced feature of each query through the deformable cross-attention as follows:

$$\mathrm{SAI}(p) = \sum_{u=1}^{U} \omega_u [\sum_{j=1}^{N} \sum_{v=1}^{N_v} \phi(\omega_f \boldsymbol{F}_i^{(t_0)}(p)) \boldsymbol{F}_{j,e}^{(t_0)}(p + \Delta p_v)], \quad (6)$$

where u is the attention head index, $\omega_{u/f}$ denotes the learnable parameters, and $\phi(\cdot)$ is the softmax function. As Figure 5 shows, we apply the filling operation to fill $\mathrm{SAI}(p)$ into $\boldsymbol{F}_i^{(t_0)}$ based on the initial position of each query and output $\boldsymbol{X}^{(t_0)}$.

Temporal Context Aggregation. Subsequently, we distill valuable target information from the temporal context and spatial semantics through the TCA component, whose input includes the memory feature $\boldsymbol{H}^{(t_0-1)}$, the ego feature $\boldsymbol{F}_i^{(t_0)}$ and the output $\boldsymbol{X}^{(t_0)}$ of the SAI component. Specifically, we first adopt the same configuration to obtain the update gate $\mathcal{G}_u^{(t_0)}$ and the reset gate $\mathcal{G}_r^{(t_0)}$:

$$\mathcal{G}_u^{(t_0)} = \sigma \cdot \boldsymbol{W}_{3\times3}([\boldsymbol{H}^{(t_0-1)};\ \boldsymbol{X}^{(t_0)}]), \quad (7)$$

where $\boldsymbol{W}_{3\times3}(\cdot)$ is the 3×3 convolution operation used for fusion, $[\cdot\,;\,\cdot]$ is the concatenation. Then, these two gates are concatenated to get the normalized weight maps by the softmax function $\phi(\cdot)$:

$$\mathcal{E}_u^{(t_0)} = \phi(\mathcal{G}_u^{(t_0)}) = \frac{exp(\mathcal{G}_u^{(t_0)})}{exp(\mathcal{G}_u^{(t_0)}) + exp(\mathcal{G}_r^{(t_0)})}. \quad (8)$$

Similarly, $\mathcal{E}_r^{(t_0)} = \phi(\mathcal{G}_r^{(t_0)})$. To improve the robustness to collaboration noises, we introduce the ego-centered characteristics in $\boldsymbol{F}_i^{(t_0)}$ into $\boldsymbol{X}^{(t_0)}$ using 1×1 convolution, which processes as follows:

$$\hat{\boldsymbol{H}}^{(t_0)} = \boldsymbol{W}_{1\times1}([\boldsymbol{X}^{(t_0)};\ \boldsymbol{F}_i^{(t_0)}]). \quad (9)$$

Finally, we filter the historical information and refine the current representation based on the two normalized weight maps. The current memory feature $\boldsymbol{H}^{(t_0)}$ is obtained as follows:

$$\boldsymbol{H}^{(t_0)} = \mathcal{E}_r^{(t_0)} \odot \boldsymbol{H}^{(t_0-1)} + \mathcal{E}_u^{(t_0)} \odot \hat{\boldsymbol{H}}^{(t_0)}, \quad (10)$$

where $\odot$ denotes the dot product operation.

3.4 Common-aware Fusion Strategy

Transmitting and fusing the common representations of collaborators is equally essential for holistic perception. We introduce a Common-Aware Fusion (CAF) strategy to aggregate the refined features $\boldsymbol{H}^{(t)} \in \mathbb{R}^{H\times W\times C}$ and the decoupled common feature maps $\boldsymbol{F}_{j,c}^{(t_0)} \in \mathbb{R}^{H\times W\times C}$, $j \in \{1, ..., N\}$ from collaborators. First, the compressed common features are restored to the previous channel dimension by a 1×1 convolutional decompressor. As shown in Figure 2 for CAF, we interpret the fusion procedure using $\boldsymbol{F}_{k,c}^{(t_0)}$ and $\boldsymbol{F}_{j,c}^{(t_0)}$ from two collaborators as examples. Concretely, the agent-wise average and max pooling mechanisms $\Phi_{a/m}(\cdot)$ are utilized to integrate the rich common semantics shared by the collaborators, denoted as $\mathcal{A}_{\{k,j\},c}^{(t_0)}, \mathcal{M}_{\{k,j\},c}^{(t_0)} = \Phi_{a/m}(\boldsymbol{F}_{\{k,j\},c}^{(t_0)})$. These features are then combined to obtain the most informative feature $\mathcal{F}_i^{(t)}$:

$$\mathcal{F}_i^{(t)} = \sigma \cdot \boldsymbol{W}_{7\times7}([\mathcal{A}_{\{k,j\},c}^{(t_0)};\ \mathcal{M}_{\{k,j\},c}^{(t_0)}]) + \boldsymbol{H}^{(t)}. \quad (11)$$

3.5 Detection Decoder and Objective Function

Upon obtaining the final fused feature map $\mathcal{F}_i^{(t)}$, two detection decoders $\{f_{dec}^r(\cdot),\ f_{dec}^c(\cdot)\}$ are applied to produce the classification and regression outputs. The regression output is obtained through $\mathcal{Y}_{i,r}^{(t)} = f_{dec}^r(\mathcal{F}_i^{(t)}) \in \mathbb{R}^{H\times W\times7}$, containing the position, size, and yaw angle of the bounding box at each location. The classification output shows the confidence value of each predefined box to be a target or background, which is $\mathcal{Y}_{i,c}^{(t)} = f_{dec}^c(\mathcal{F}_i^{(t)}) \in \mathbb{R}^{H\times W\times2}$. For objective optimization, we adopt the smooth absolute error loss $\mathcal{L}_{reg}$ for regressing the bounding boxes and the focal loss [17] $\mathcal{L}_{cla}$ for classification. Combining the decoupling losses $\mathcal{L}_{exc},\ \mathcal{L}_{com}$ and the task losses $\mathcal{L}_{reg},\ \mathcal{L}_{cla}$, the objective function is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{reg} + \mathcal{L}_{cla} + \alpha \cdot \mathcal{L}_{exc} + \beta \cdot \mathcal{L}_{com}, \quad (12)$$

where α and β are trade-off parameters.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Datasets. To evaluate the performance of our framework on the LiDAR-based collaborative 3D object detection task, we conduct extensive experiments on three public multi-agent datasets, including DAIR-V2X [52], V2XSet [40], and OPV2V [41]. **DAIR-V2X** [52] is the first large-scale real-world dataset for supporting collaborative perception, containing the labeled LiDAR point clouds of a vehicle and an infrastructure. It includes 100 autonomous driving scenes and 18,000 data samples, where training/validation/testing sets are split in a 5:2:3 ratio. **V2XSet** [40] is a large-scale simulation dataset for V2X perception. 73 representative scenarios are included in this dataset, where each scenario lasts 25 seconds and contains 2 to 5 connected agents. The training/validation/testing sets are 6,694/1,920/2,833 frames, respectively. **OPV2V** [41] is a vehicle-to-vehicle collaborative perception dataset co-simulated by CARLA [4] and OpenCDA [36]. This dataset contains 11,464 LiDAR point cloud frames with 3D annotations, of which training/validation/testing sets contain 6,764, 1,981, and 2,719 frames, respectively.

Evaluation Metrics. To evaluate the 3D object detection performance of the proposed framework, we employ the Average Precision (AP) at Intersection-over-Union (IoU) thresholds of 0.5 and 0.7 as precision metrics. Moreover, we leverage the same calculation format as [6] to count the communication volume by byte in the log scale with base 2.

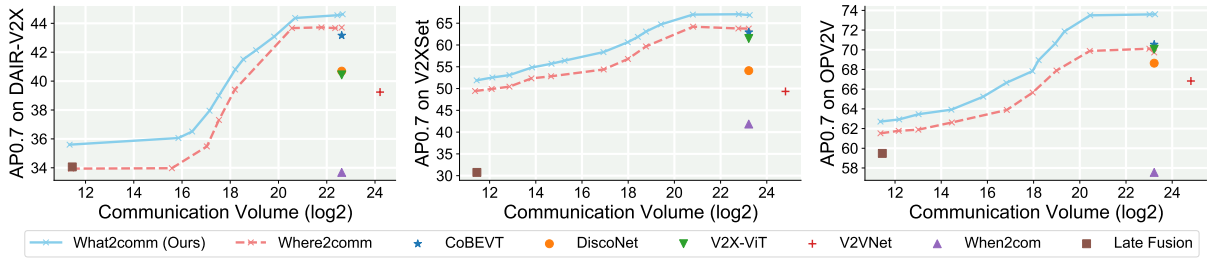


Figure 6: Collaborative perception performance comparison of What2comm and Where2comm [6] on the DAIR-V2X, V2XSet and OPV2V with varying communication volumes.

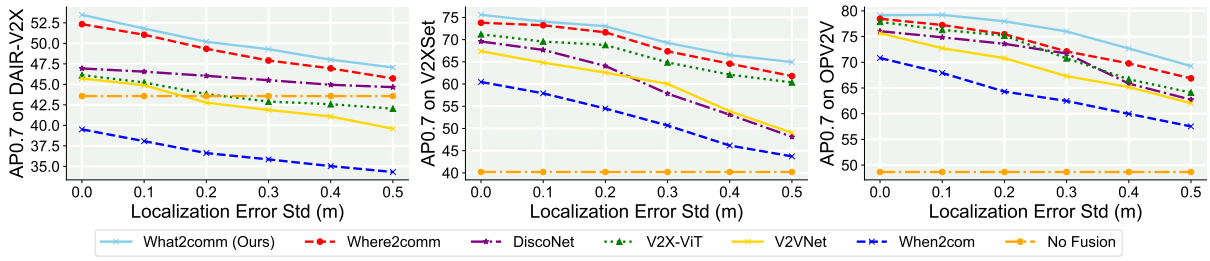


Figure 7: Robustness to the localization error on the DAIR-V2X, V2XSet, and OPV2V datasets.

4.2 Implementation Details

Our models are built based on Pytorch toolbox [25]. We train them on 4 Tesla V100 GPUs using the Adam optimizer [9]. The initial learning rate is 2e-3 and decays by an exponential factor of 0.1 every 15 epochs. The batch sizes on the DAIR-V2X [52], V2XSet [40], and OPV2V [41] datasets are $\{5, 3, 3\}$, and the training epochs are $\{30, 40, 40\}$. We set the voxel resolution of the feature encoder $f_{enc}(\cdot)$ on height and width to 0.4 m. The importance generator $f_{gen}(\cdot)$ is implemented based on the detection decoder in [11], and the pixel number N_r are set as 2048. In the STCM component, the attention head is 8, and the keypoint number N_v is 15. We build the common and exclusive encoders and detection decoders using 1×1 convolutional layers. The trade-off parameters α and β are 2e-3 and 1e-2, respectively. To simulate the default collaboration noises, we consider that the localization and heading errors of the collaborators follow a Gaussian distribution with standard deviations of 0.2 m and 0.2°, respectively. The transmission delay between the ego agent and collaborators is set to 100 ms. Below, we present the results from the validation set of DAIR-V2X and the testing sets of V2XSet and OPV2V.

4.3 Quantitative Evaluation

Detection Performance Evaluation. From Table 1, we compare the object detection performance of What2comm and the existing methods across the three datasets. No Fusion considers only the point cloud of the ego agent and performs detection via the PointPillar detector [11]. Late Fusion allows agents to share the predicted 3D boxes and produce outputs using non-maximum suppression. Moreover, the previous state-of-the-art (SOTA) models are comprehensively considered, including When2com [22], V2VNet [32], V2X-ViT [40], DiscoNet [15], Where2comm [6], and CoBEVT [38]

Table 1: Performance comparison on the DAIR-V2X, V2XSet, and OPV2V datasets. The results are reported in AP@0.5/0.7.

Model	DAIR-V2X	V2XSet	OPV2V
	AP@0.5/0.7	AP@0.5/0.7	AP@0.5/0.7
No Fusion	50.03/43.57	60.60/40.20	68.71/48.66
Late Fusion	48.93/34.06	54.92/30.75	79.62/59.48
When2com [22]	48.20/33.68	67.41/41.85	74.11/57.55
V2VNet [32]	53.46/39.24	79.17/49.34	80.31/66.83
V2X-ViT [40]	53.08/40.43	83.63/61.49	84.65/70.06
DiscoNet [15]	52.67/40.69	79.82/54.11	84.72/68.64
Where2comm [6]	59.52/43.71	83.17/63.77	85.16/69.73
CoBEVT [38]	58.38/43.16	82.77/62.93	85.49/70.54
What2comm (Ours)	**60.81/44.63**	**84.59/66.86**	**86.83/73.61**

(mentioned in Section 2.1). Intuitively, the proposed What2comm greatly outperforms the other methods under the default noise settings, demonstrating the robustness of our model against collaboration noises. In particular, compared to the SOTA performance of AP@0.7, What2comm improves by 2.1% on the real-world dataset (DAIR-V2X [52]), 4.8% on V2XSet [40], and 4.4% on OPV2V [41]. The reasonable explanations include: (**i**) The decoupled exclusive and common representations facilitate perspective completion and eliminate data discrepancies, respectively. (**ii**) The proposed STCM module alleviates the feature map misalignment caused by noises.

Comparison of Communication Volume. Figure 6 shows the performance comparison results under various bandwidth consumptions. Concretely, the blue and red curves indicate the evaluated performance of our What2comm and Where2comm [6] under varying communication volumes, respectively. We have the following observations: (**i**) What2comm achieves the same detection performance as the SOTA model with less communication volume on

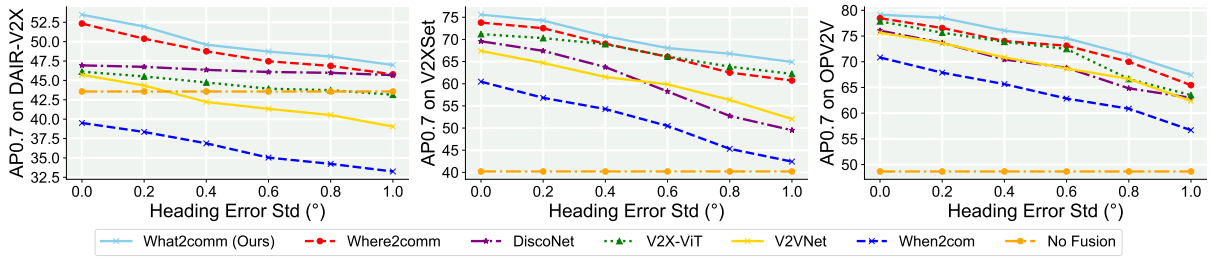


Figure 8: Robustness to the heading error on the DAIR-V2X, V2XSet, and OPV2V datasets.

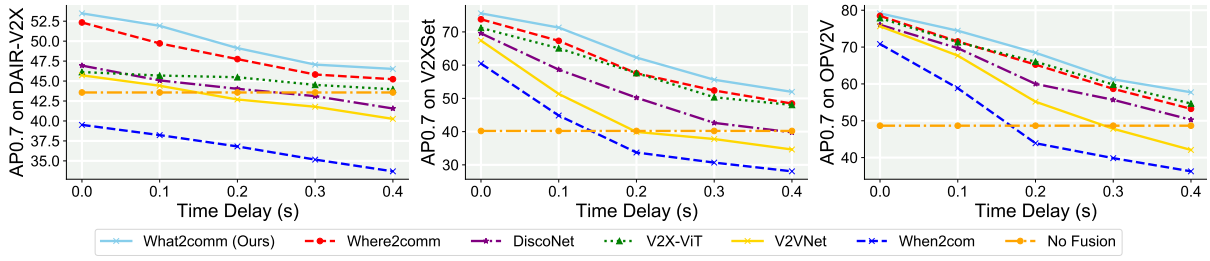


Figure 9: Robustness to the time delay on the DAIR-V2X, V2XSet, and OPV2V datasets.

both real-world and simulated datasets. (**ii**) Compared with the spatial filtering-based method Where2comm [6], our model achieves a superior performance-bandwidth tradeoff under all bandwidth choices via feature decoupling. The noteworthy improvement reveals the effectiveness of our communication mechanism.

Robustness to Localization and Heading Errors. Here, we adopt the same localization noise settings as [40] and conduct extensive experiments on the three datasets to evaluate the sensitivity of existing methods to localization and heading errors. As Figures 7&8 show, the noises are sampled from Gaussian distributions with standard deviations of $[0, 0.5]$ m and $[0°, 1.0°]$, respectively. Especially, the performance of all feature-level fusion-based methods inevitably decreases due to feature misalignment with increasing localization and heading errors. Some of them are even weaker than No Fusion under massive errors. In comparison, although our method (blue curve) shows a downward trend, it consistently maintains higher detection performance than existing SOTA methods. The plausible reasons are: (**i**) Our SAI component facilitates capturing the perceptually critical information from the decoupled exclusive representations through the deformable cross-attention. (**ii**) The proposed component TCA overcomes the collaboration noises by adaptively integrating temporal context information and ego representation based on their perception contributions.

Robustness to Time Delay. In addition to the localization noises, the temporal asynchrony due to time delay will also degrade the performance of the existing collaborative perception models. As shown in Figure 9, the detection performance of all collaboration methods decreases with increasing time delay (ranging from 0 to 400 ms). For instance, the AP@0.7 of DiscoNet [15], When2com [22], and V2VNet [32] on DAIR-V2X decrease significantly and are worse than the baseline No Fusion when the time delay exceeds 300 ms. Contrastly, What2comm improves the SOTA performance of the three datasets under all time delay levels and maintains high detection accuracy even under a severe delay (400 ms). The comparison results prove that What2comm is more robust to time delay than previous SOTA methods since it captures informative historical context cues and introduces ego-centered characteristics.

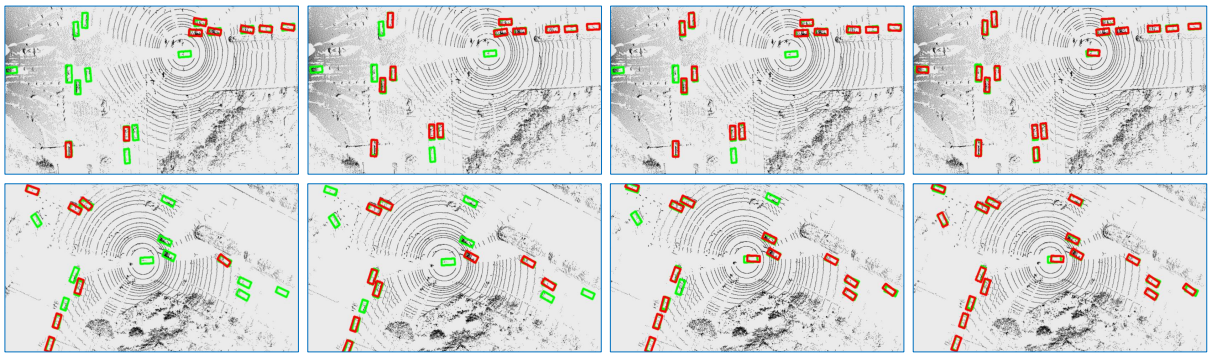
Table 2: Ablation studies of the proposed core components on the three datasets. SAI: spatial attention integration; TCA: temporal context aggregation; SC: specificity constraint; CC: consistency constraint; CAF: common-aware fusion strategy.

SAI	TCA	SC	CC	CAF	DAIR-V2X AP@0.5/0.7	V2XSet AP@0.5/0.7	OPV2V AP@0.5/0.7
					53.47/38.78	76.90/52.85	77.44/64.97
✔					57.06/42.76	79.44/63.29	81.88/68.26
✔	✔				58.30/43.07	81.25/64.11	83.29/69.94
✔	✔	✔			60.15/43.98	83.87/66.05	85.66/72.75
✔	✔	✔	✔		60.38/44.46	84.31/66.53	86.29/73.27
✔	✔	✔	✔	✔	**60.81/44.63**	**84.59/66.86**	**86.83/73.61**

4.4 Ablation Studies

We perform thorough ablation studies on three datasets to verify the necessity of different mechanisms and design philosophies in What2comm. Tables 2 and 3 show the following observations.

Importance of Core Components. The contribution of each component is systematically investigated in Table 2. We first introduce a baseline version as a gain reference, where the collaborator-shared features and ego-centered temporal information are fused by a position-wise maximum mechanism and a 1×1 convolutional operation, respectively. In this case, we progressively add (1) spatial attention integration (SAI), (2) temporal context aggregation (TCA), (3) specificity constraint (SC), (4) consistency constraint (CC), and (5) common-aware fusion (CAF) and present the corresponding



(a) No Fusion (b) CoBEVT (c) Where2comm (d) What2comm (ours)

Figure 10: Qualitative comparison in real-world scenarios from the DAIR-V2X dataset. Green and red boxes denote ground truth and detection, respectively. Compared to the previous SOTA models, our method achieves more accurate detection results.

Table 3: Ablation study results of candidate designs and strategies on the three datasets. "w/" means the with.

Designs/Strategies	DAIR-V2X	V2XSet	OPV2V
	AP@0.5/0.7	AP@0.5/0.7	AP@0.5/0.7
Full Model	**60.81/44.63**	**84.59/66.86**	**86.83/73.61**
Effect of Keypoint Number			
12 Keypoints	60.09/44.12	83.88/65.97	86.18/72.21
15 Keypoints (Default)	60.81/44.63	84.59/66.86	86.83/73.61
18 Keypoints	60.47/44.38	84.21/66.32	86.25/72.74
Rationality of Fusion Strategies			
w/ Summation Fusion	59.31/43.29	82.68/64.06	83.92/69.93
w/ Average Fusion	60.20/43.73	83.79/65.41	85.48/72.44
w/ Maximum Fusion	60.42/44.02	83.22/66.03	85.02/71.81

detection precision. According to the result variations among the three datasets, we find that all components are beneficial to the performance gains. Also, SAI and SC have the most valuable contributions due to the significant improvements they bring to all datasets. For example, SAI increases AP@0.7 by 3.98%, 10.44%, and 3.29% on the DAIR-V2X, V2XSet, and OPV2V, respectively.

Effect of Keypoint Number. The choice of the keypoint number in spatial attention integration plays an essential role in the information fusion among heterogeneous agents. In the upper part of Table 3, we evaluate the effect of the keypoint number on detection performance. Specifically, we empirically sample the appropriate number of keypoints and find that better detection performance is achieved when the number is 15. This phenomenon implies that setting a reasonable keypoint number facilitates our component to effectively capture spatially critical semantic information and potentially mitigate the interference of localization errors.

Rationality of Fusion Strategies. Here, we justify the proposed common-aware fusion strategy by introducing several heuristic fusion mechanisms. Concretely, Summation Fusion is a straightforward pixel-wise addition operation. Average and Maximum Fusion select the average and maximum values of the elements at the corresponding positions among feature maps and convert them into a fused feature map. As shown in the bottom part of Table 3, the above candidate mechanisms fail to tackle the multi-source feature fusion challenge due to the sub-optimal performance achieved. In contrast, our feature fusion strategy brings better detection results with the advantage of effectively filtering redundant information from decoupled common features.

4.5 Qualitative Evaluation

In order to qualitatively compare the perception performance of different methods, Figure 10 provides visualizations of the detection results of No Fusion, CoBEVT [38], Where2comm [6], and What2comm in two scenes under default noise settings. Intuitively, our model achieves more comprehensive and accurate detection results than the previous SOTA methods and No Fusion. On the one hand, What2comm is able to produce more bounding boxes corresponding to the ground truth. The reason is that What2comm extends the perceptual range and completes the occluded areas based on the proposed efficient communication mechanism. On the other hand, previous SOTA approaches fail to predict well-aligned bounding boxes for fast-moving targets due to collaboration noises. In contrast, What2comm aggregates spatially complementary information by sampling keypoints and adaptively integrates spatial-temporal semantics, leading to a more informative fused visual representation and a more holistic perception.

5 CONCLUSION

In this paper, we present What2comm, a novel multi-agent collaboration framework that seeks a trade-off between perception performance and communication bandwidth in an end-to-end manner. What2comm achieves efficient information transmission across agents via the feature decoupling. Subsequently, a spatio-temporal collaboration module and a feature fusion strategy are sequentially proposed to aggregate historical cues from the ego agent and to fuse decoupled features from collaborators. Extensive experiments on real-world and simulated datasets show the effectiveness of What2comm and the rationality of all its vital components.

ACKNOWLEDGMENTS

This work is supported by the Shanghai Key Research Laboratory of NSAI, the National Natural Science Foundation of China (Grant No. 62250410368), and the Nanjing First Automobile Works Grant.

REFERENCES

[1] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 29.

[2] Runjian Chen, Yao Mu, Runsen Xu, Wenqi Shao, Chenhan Jiang, Hang Xu, Zhenguo Li, and Ping Luo. 2022. CO^ 3: Cooperative Unsupervised 3D Representation Learning for Autonomous Driving. *arXiv preprint arXiv:2206.04028* (2022).

[3] Zhaoyu Chen, Bo Li, Shuang Wu, Jianghe Xu, Shouhong Ding, and Wenqiang Zhang. 2022. Shape matters: deformable patch attack. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 529–548.

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

[5] Yangtao Du, Dingkang Yang, Peng Zhai, Mingchen Li, and Lihua Zhang. 2021. Learning Associative Representation for Facial Expression Recognition. In *IEEE International Conference on Image Processing (ICIP)*. 889–893.

[6] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. 2022. Where2comm: Communication-Efficient Collaborative Perception via Spatial Confidence Maps. In *Advances in Neural Information Processing Systems (NIPS)*.

[7] Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, and Yanfeng Wang. 2023. Collaboration Helps Camera Overtake LiDAR in 3D Detection. *arXiv preprint arXiv:2303.13560* (2023).

[8] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*. PMLR, 2649–2658.

[9] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

[10] Haopeng Kuang, Dingkang Yang, Shunli Wang, Xiaoying Wang, and Lihua Zhang. 2023. Towards Simultaneous Segmentation Of Liver Tumors And Intrahepatic Vessels Via Cross-Attention Mechanism. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[11] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12697–12705.

[12] Yuxuan Lei, Dingkang Yang, Mingcheng Li, Shunli Wang, Jiawei Chen, and Lihua Zhang. 2023. Text-oriented Modality Reinforcement Network for Multimodal Sentiment Analysis from Unaligned Multimodal Sequences. *arXiv preprint arXiv:2307.13205* (2023).

[13] Zixing Lei, Shunli Ren, Yue Hu, Wenjun Zhang, and Siheng Chen. 2022. Latency-aware collaborative perception. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 316–332.

[14] Jinlong Li, Runsheng Xu, Xinyu Liu, Jin Ma, Zicheng Chi, Jiaqi Ma, and Hongkai Yu. 2023. Learning for vehicle-to-vehicle cooperative perception under lossy communication. *IEEE Transactions on Intelligent Vehicles* (2023).

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

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2117–2125.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[18] Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Zhao Zhile Yang, Dingkang, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, and Gan Zhongxue. 2023. Improving Generalization in Visual Reinforcement Learning via Conflict-aware Gradient Agreement Augmentation. *arXiv preprint arXiv:2308.01194* (2023).

[19] Yang Liu, Jing Liu, Kun Yang, Bobo Ju, Siao Liu, Yuzheng Wang, Dingkang Yang, Peng Sun, and Liang Song. 2023. AMP-Net: Appearance-Motion Prototype Network Assisted Automatic Video Anomaly Detection System. *IEEE Transactions on Industrial Informatics* (2023), 1–13. https://doi.org/10.1109/TII.2023.3298476

[20] Yang Liu, Jing Liu, Mengyang Zhao, Dingkang Yang, Xiaoguang Zhu, and Liang Song. 2022. Learning appearance-motion normality for video anomaly detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.

[21] Yang Liu, Dingkang Yang, Yan Wang, Jing Liu, and Liang Song. 2023. Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models. *arXiv preprint arXiv:2302.05087* (2023).

[22] Yen-Cheng Liu, Junjiao Tian, Nathaniel Glaser, and Zsolt Kira. 2020. When2com: Multi-agent perception via communication graph grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4106–4115.

[23] Guiyang Luo, Hui Zhang, Quan Yuan, and Jinglin Li. 2022. Complementarity-Enhanced and Redundancy-Minimized Collaboration Network for Multi-agent Perception. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 3578–3586.

[24] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning (ICML)*. PMLR, 2642–2651.

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NIPS)* 32 (2019).

[26] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 652–660.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 234–241.

[28] Nicholas Vadivelu, Mengye Ren, James Tu, Jingkang Wang, and Raquel Urtasun. 2021. Learning to communicate and correct pose errors. In *Conference on Robot Learning (CoRL)*. PMLR, 1195–1210.

[29] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Yan Wang, and Wenqiang Zhang. 2022. Boosting the Transferability of Adversarial Attacks with Global Momentum Initialization. *arXiv preprint arXiv:2211.11236* (2022).

[30] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. 2021. Tsa-net: Tube self-attention network for action quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia (ACM MM)*. 4902–4910.

[31] Tianhang Wang, Guang Chen, Kai Chen, Zhengfa Liu, Bo Zhang, Alois Knoll, and Changjun Jiang. 2023. UMC: A Unified Bandwidth-efficient and Multi-resolution based Collaborative Perception Framework. *arXiv preprint arXiv:2303.12400* (2023).

[32] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 605–621.

[33] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. 2019. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 9005–9012.

[34] Hao Xiang, Runsheng Xu, Xin Xia, Zhaoliang Zheng, Bolei Zhou, and Jiaqi Ma. 2023. V2xp-asg: Generating adversarial scenes for vehicle-to-everything perception. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3584–3591.

[35] Runsheng Xu, Weizhe Chen, Hao Xiang, Xin Xia, Lantao Liu, and Jiaqi Ma. 2023. Model-agnostic multi-agent perception framework. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1471–1478.

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

[37] Runsheng Xu, Jinlong Li, Xiaoyu Dong, Hongkai Yu, and Jiaqi Ma. 2022. Bridging the domain gap for multi-agent perception. *arXiv preprint arXiv:2210.08451* (2022).

[38] Runsheng Xu, Zhengzhong Tu, Hao Xiang, Wei Shao, Bolei Zhou, and Jiaqi Ma. 2022. CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers. In *Conference on Robot Learning (CoRL)*.

[39] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13712–13722.

[40] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. 2022. V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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

[42] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. 2023. Context De-Confounded Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19005–19015.

[43] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 1642–1651.

[44] Dingkang Yang, Shuai Huang, Yang Liu, and Lihua Zhang. 2022. Contextual and Cross-Modal Interaction for Multi-Modal Speech Emotion Recognition. *IEEE Signal Processing Letters* 29 (2022), 2093–2097.

[45] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. 2022. Emotion Recognition for Multiple Context

Awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Vol. 13697. Springer, 144–162.

[46] Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. 2023. AIDE: A Vision-Driven Multi-View, Multi-Modal, Multi-Tasking Dataset for Assistive Driving Perception. *arXiv preprint arXiv:2307.13933* (2023).

[47] Dingkang Yang, Haopeng Kuang, Shuai Huang, and Lihua Zhang. 2022. Learning Modality-Specific and -Agnostic Representations for Asynchronous Multimodal Language Sequences. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*. 1708–1717.

[48] Dingkang Yang, Yang Liu, Can Huang, Mingcheng Li, Xiao Zhao, Yuzheng Wang, Kun Yang, Yan Wang, Peng Zhai, and Lihua Zhang. 2023. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems* (2023), 110370.

[49] Kun Yang, Jing Liu, Dingkang Yang, Hanqi Wang, Peng Sun, Yanni Zhang, Yan Liu, and Liang Song. 2023. A novel efficient Multi-view traffic-related object detection framework. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[50] Kun Yang, Peng Sun, Jieyu Lin, Azzedine Boukerche, and Liang Song. 2022. A novel distributed task scheduling framework for supporting vehicular edge intelligence. In *IEEE International Conference on Distributed Computing Systems (ICDCS)*. 972–982.

[51] Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, and Liang Song. 2023. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. *arXiv preprint arXiv:2307.13929* (2023).

[52] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21361–21370.

[53] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Jirui Yuan, Ping Luo, and Zaiqing Nie. 2023. Vehicle-Infrastructure Cooperative 3D Object Detection via Feature Flow Prediction. *arXiv preprint arXiv:2303.10552* (2023).

[54] Xiaoding Yuan, Adam Kortylewski, Yihong Sun, and Alan Yuille. 2021. Robust instance segmentation through reasoning about multi-object occlusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11141–11150.

[55] Zhenxun Yuan, Xiao Song, Lei Bai, Zhe Wang, and Wanli Ouyang. 2021. Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 4 (2021), 2068–2078.

[56] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811* (2017).

[57] Kaihao Zhang, Rongqing Li, Yanjiang Yu, Wenhan Luo, and Changsheng Li. 2021. Deep dense multi-scale network for snow removal using semantic and depth priors. *IEEE Transactions on Image Processing* 30 (2021), 7419–7431.

[58] Xiao Zhao, Liuzhen Su, Xukun Zhang, Dingkang Yang, Mingyang Sun, Shunli Wang, Peng Zhai, and Lihua Zhang. 2023. D-CONFORMER: Deformable Sparse Transformer Augmented Convolution for Voxel-Based 3D Object Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[59] Ruichao Zhu, Jiafu Wang, Tianshuo Qiu, Dingkang Yang, Bo Feng, Zuntian Chu, Tonghao Liu, Yajuan Han, Hongya Chen, and Shaobo Qu. 2023. Direct field-to-pattern monolithic design of holographic metasurface via residual encoder-decoder convolutional neural network. *Opto-Electronic Advances* (2023), 220148–1.

[60] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *International Conference on Learning Representations (ICLR)*.