V2VFormer++: Multi-Modal Vehicle-to-Vehicle Cooperative Perception via Global-Local Transformer

Hongbo Yin, Daxin Tian, *Senior Member, IEEE*, Chunmian Lin, Xuting Duan, *Member, IEEE*, Jianshan Zhou, Dezong Zhao, *Senior Member, IEEE*, and Dongpu Cao

***Abstract*— Multi-vehicle cooperative perception has recently emerged for facilitating long-range and large-scale perception ability of connected automated vehicles (CAVs). Nonetheless, enormous efforts formulate collaborative perception as LiDAR-only 3D detection paradigm, neglecting the significance and complementary of dense image. In this work, we construct the first multi-modal vehicle-to-vehicle cooperative perception framework dubbed as V2VFormer++, where individual camera-LiDAR representation is incorporated with dynamic channel fusion (DCF) at bird's-eye-view (BEV) space and ego-centric BEV maps from adjacent vehicles are aggregated by global-local transformer module. Specifically, channel-token mixer (CTM) with MLP design is developed to capture global response among neighboring CAVs, and position-aware fusion (PAF) further investigate the spatial correlation between each ego-networked map in a local perspective. In this manner, we could strategically determine which CAVs are desirable for collaboration and how to aggregate the foremost information from them. Quantitative and qualitative experiments are conducted on both publicly-available OPV2V and V2X-Sim 2.0 benchmarks, and our proposed V2VFormer++ reports the state-of-the-art cooperative perception performance, demonstrating its effectiveness and advancement. Moreover, ablation study and visualization analysis further suggest the strong robustness against diverse disturbances from real-world scenarios.**

***Index Terms*— Vehicle-to-vehicle (V2V) cooperative perception, multi-modal fused perception, autonomous driving, transformer, 3D object detection, intelligent transportation systems.**

I. Introduction

AS THE cutting-edge technology, autonomous driving is regarded as the trend of intelligent transportation system (ITS), that provides a promising solution to troublesome problems including traffic congestion, collision, and emission pollution [1]. With the development of deep learning and computer vision, environmental perception as the essential component of self-driving system, has also made great progress on such object detection [2], [3], [4], [5] and segmentation [6], [7] tasks, receiving substantial performance improvement both on accuracy and efficiency. Due to the complex traffic scenarios and varying physical conditions, it is difficult to ensure the robust and safe sensing performance purely depended on ego-view information. Therefore, how to exploit and aggregate multi-source information to enhance the perception ability is the hot-spot issue both in academia and industry.

Vehicle-to-vehicle (V2V) cooperative perception has recently emerged based on information fusion and data sharing, that strategically incorporates multi-view surroundings from neighboring connected automated vehicles (CAVs) via low-latency vehicular communication [8]. In this way, several perception challenges occurred in various driving scenario, i.e., bind spot, beyond line-of-sight, occlusion, etc., could be significantly alleviated, simultaneously enabling self-driving car with long-range and large-scale perception ability as shown in Fig. 1. According to different collaborative strategies, current works [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] on multi-agent perception can be divided into three folds. On one hand, early fusion [9] transmits raw sensor data from each vehicle into a targeted one, however, it fails to meet the requirement of real-time system due to the unaffordable computational overhead and communication bandwidth. On the other hand, late fusion [10], [11] straightforwardly re-weights the individual detections via mathematical operations (i.e., sum and average) or attention mechanism, greatly improving the running speed. Whereas, false positives in different vehicles could be amplified in this way, and the accumulated spatial displacement would damage multi-agent collaborative performance step-by-step. Intermediate feature collaboration [12], [13], [14], [15], [16], [17], [18] has gained increasing popularity due to its better trade-off between accuracy and speed, that projects compact feature representation (e.g., BEV map) from CAVs into a unified coordinate for comprehensively understanding the traffic scenario in a global view. Given multiple BEV features, graph-based methods [12], [13], [14], [15], [16], [17] create

Manuscript received 29 May 2023; revised 4 August 2023; accepted 7 September 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC3803700; and in part by the National Natural Science Foundation of China under Grant U20A20155, Grant 62173012, and Grant 52202391. The Associate Editor for this article was M. Yang. *(Corresponding author: Chunmian Lin.)*

Hongbo Yin, Daxin Tian, Chunmian Lin, Xuting Duan, and Jianshan Zhou are with the State Key Laboratory of Intelligent Transportation System, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China (e-mail: cmlin@buaa.edu.cn).

Dezong Zhao is with the James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K.

Dongpu Cao is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada.

Digital Object Identifier 10.1109/TITS.2023.3314919

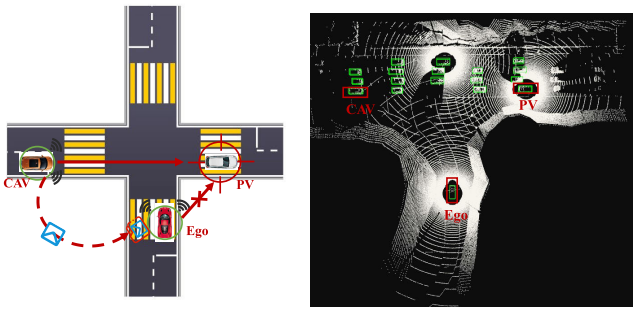


Fig. 1. The advantage of cooperative perception. **Left.** The car fails to perceive the potential threat of public vehicle (PV) in advance when driving at the intersection. With the help of connected automated vehicle (CAV), holistic-view information could be transmitted to the ego-vehicle (Ego) for circumventing traffic conflict in the blind-spot area. **Right.** The detection result is visualized in the bird's-eye-view (BEV) space.

a weighted collaboration diagram where each node denotes a single vehicle with their real-time pose information, and pair-wise edge defines the relationship between adjacent vehicles via spatial weight matrix. Moreover, transformer-based algorithms perform self-attention or cross-attention operation on the encoded sequence to capture both local and global dependencies among CAVs. As an alternative, keypoint representation is efficiently selected by Farthest Point Sampling (FPS) for highlighting significant point, which would be further preserved only if it exists in candidate proposal [18]. In conclusion, the pipeline of these works is to combine inter-vehicle representations for feature enhancement at the ego coordinate, while it easily suffers from feature ambiguity and semantic deficiencies because of the sparsity and unevenness of LiDAR point. More importantly, previous collaboration strategies explicitly construct spatial feature relation after projecting into a reference coordinate, but it is non-trivial to investigate the channel interactions across CAVs in the global.

In this paper, we cast vehicle-to-vehicle perception into 3D object detection task, and develop multi-modal vehicle-to-vehicle cooperative perception framework termed as V2VFormer++. For each CAV, a camera-LiDAR paradigm is proposed for overcoming the drawback of LiDAR-only detection, which compensates accurate geometry with dense context (i.e., texture, profile, etc.) for versatile surrounding description. To obtain the expressive representation in the unified plane, we project two heterogeneous modalities into the birds'-eye view (BEV) space via view transformation, and further design a simple yet efficient multi-modal fusion module dynamic channel fusion (DCF) for pixel-point correspondence aggregation in an adaptative manner. In this way, abundant semantic attribute at each ego-centric perspective could be adequately exploited with marginal computational budget. As for vehicle-to-vehicle perception, a novel global-local transformer strategy is proposed to aggregate intermediate features from CAVs. Specifically, we firstly adopt the channel-token mixer (CTM) with MLP design to calculate the global response among different vehicles, and thus each ego-networked pair is matched according to correlation score. To concern more about the region of interest (RoI), position-aware fusion (PAF) is introduced for attending to the informative area across all vehicles, and pixel-wise feature semantics in the local is also explored with a self-attention transformer. Finally, we perform comprehensive empirical studies on both OPV2V [19] and V2X-Sim 2.0 [20] datasets, the proposed V2VFormer++ achieves the state-of-the-art collaborative perception accuracy, which overperforms the counterparts (e.g., multi-vehicle single-modal and multi-vehicle multi-modal) by a substantial margin. Moreover, ablation analysis on diverse configuration and scenarios further suggests its robustness and generality against real-world interruptions.

In summary, the contributions of this work are mainly described as follows:

[1] We propose V2VFormer++, the first multi-modal vehicle-to-vehicle cooperative perception framework that consumes heterogeneous modalities from separate vehicle for enhancing multi-agent collaboration performance.

[2] Dynamic channel fusion (DCF) module is designed for correspondence aggregation from camera and LiDAR BEV maps in an adaptative manner.

[3] Global-local transformer collaboration is an intermediate fusion strategy where channel-token mixer (CTM) is developed for capturing global response among CAVs and position-aware fusion (PAF) module is utilized to explore spatial semantics in a local perspective.

[4] Without bells and whistles, our V2VFormer++ reports the state-of-the-art cooperative detection performace on both OPV2V [19] and V2X-Sim 2.0 [20] benchmarks, which outperforms all alternatives over a remarkable margin. Besides, ablation study and visualization result further demonstrate its robustness againse varied disturbances from real-world scenarios.

II. Related Works

This section aims to review the related studies on LiDAR-based and camera-LiDAR 3D object detection, as well as vehicle-to-vehicle cooperative perception.

A. LiDAR-Based 3D Detection

According to various data formats, LiDAR-based 3D detection can be broadly divided into three categories: point-based, voxel-based and hybrid representations. The pipeline of point-based algorithm directly consumes the raw LiDAR data with PointNet [21] and PointNet++ [22] architectures for reliable geometrical feature extraction, the former of which adopts set abstraction (SA) operator to aggregate point-wise representation and utilizes transform network (T-Net) for feature alignment in both input and feature levels, while the latter of which further learning both local and global contexts from point cloud via hierarchical sampling method. For 3D detection task, 3DSSD [23] simultaneously introduce distance (D-) and feature (F-) farthest point sampling (FPS) strategies to handle the sparsity of point representation, and conduct object localization and classification via an SSD (single-stage detector)-like architecture. To better distinguish foreground point from the background, CenterPoint [24] extracts keypoint feature to predict 3D bounding boxes from the center points

of objects, and IA-SSD [24] leverages instance-aware features with SSD architecture for 3D object detection. Voxel-based detector is an efficient paradigm where 3D point space is firstly discretized into regular grids, and convolutional network is then introduced to process the fine-grained feature encoded within each voxel. As the pioneering work, VoxelNet [25] designs stacked voxel feature encoding (VFE) layer to point-wise information extraction, and 3D convolution is then utilized for intermediate feature aggregation in the local. To accelerate the inference speed, SECOND [26] develop 3D sparse convolution for high-efficiency voxel feature encoding, while PointPillars [27] collapses point cloud into a 2D representation and uses sparse convolutional backbone instead. Motivated by the idea of 2D Faster RCNN [28], Deng et al. [29] propose a two-stage 3D detection framework named as Voxel RCNN with better trade-off between accuracy and efficiency, that firstly generates coarse 3D candidate proposals and performs box refinement via voxel RoI pooling layer in the second stage. Moreover, CAGroup3D [30] explores full convolution 3D pooling to enhance the backbone feature within each proposal box, pursuing for ultimate detection performance. Studies on incorporating point-wise with voxel-wise features for 3D object detection have recently been a hot-spot issue. STD [31] follows a sparse-to-dense detection paradigm that obtains accurate proposals from the raw point with novel spherical anchor, and generates compact representation from sparse point expression via pointspool. PV-RCNN [32] summarizes the 3D scene into a set of keypoint with voxel set abstraction module, and abstracts proposal-specific feature into a dense grid by RoI grid pooling. Besides, PV-RCNN++ [33] introduces a position-sensitive fusion module for feature enhancement both on point cloud and voxel grids. Part-A^2 Net [34] consists of part-aware and part-aggregation stage, the former of which aims at high-quality proposal generation with intra-object part location, while the latter of which conducts box refinement according to the spatial location relationship after pooling. SE-SSD [35] adopts a pair of teacher and student detectors with an effective IoU-based matching strategy and consistency ODIoU loss for performance boost. Moreover, Noh et al. [36] proposes a new HVPR architecture that integrates point-based and voxel-based features into a single 3D representation, and designs attentive multi-scale feature module to learn scale-aware information from sparse and irregular point patterns. In this paper, we adopt PointPillars as the single-vehicle LiDAR backbone for the trade-off between efficiency and accuracy.

B. Camera-LiDAR 3D Object Detection

Camera-LiDAR fused perception [5] has demonstrated its meliority and drawn broad attention for 3D detection recently, which compensates the sparsity, uncertainty, and semantic fragmentation of lidar-only methods. Without the sophisticated process for pseudo-LiDAR generation, Pointpainting [37] designs a sequential-based fusion mechanisms that firstly decorates raw point cloud with the pixel-wise semantic score produced by image segmentation network and then put them into any LiDAR-only pipeline. 3D-CVF [38] projects a dense camera voxel onto the BEV plane, and concatenates each modality through an adaptive gated attention map. Furthermore, Chen et al. [39] establishes pixel-voxel view association via a learnable perspective alignment rather than an inherent projection matrix, being flexibly desirable for the consistency of heterogeneous representation. Yang et al. [40] introduces a novel modality-specific encoder-decoder structure with the throughout bilateral cross-attention span coordinate to reserve the utmost intra-characteristic in an unmixing way. To bridge the information gap between image and LiDAR, MVP [41] lifts each pixel into 3D virtual point for gathering geometrical structure, while Jiao et al. [42] proposes a multi-depth unprotection (MDU) block to compensate the depth blur and the mismatch of multi-granularity geometric for more pronounced detection. Recently, Transfusion [43] is the first attempt to introduce the transformer into camera-LiDAR 3D detection due to its superiority in long-range dependency modeling. It applies two sequential decoder layers to softly associate object query with the coarse LiDAR and fine-gained image features on BEV plane, enhancing the perception performance stage-by-stage. Similarly, UVTR [44] extends image-specific space into the voxel by transformer-based decoder and probability depth distribution, and further performs cross-attention feature interaction via knowledge transfer. BEVFusion [45] converts multi-modal streams into a canonical coordinate, and adopt a dynamic fusion strategy to prevent the failure case from LiDAR malfunction. In this paper, we aim at a simple and grace pixel-point fusion paradigm where heterogeneous feature could be transformed into a unified representation, and two BEV maps are projected onto the height-agnostic ego plane in a self-adaptation aggregation.

C. Vehicle-to-Vehicle Cooperative Perception

Vehicle-to-vehicle (V2V) collaborative perception has recently emerged with advanced vehicular communication and information fusion, and thus provides an effective solution to alleviate the beyond-line-of-sight and blind-spot challenges caused by single-agent detector. In general, this pipeline incorporates multi-view surrounding sensory data from connected automated vehicles (CAVs) with the ego-centric observations to facilitate the global perception ability, and according to different cooperative phase, prior works could be mainly divided into early, immediate, and late collaboration. Cooper [9] primarily shares multi-resolution LiDAR point, and projects own sparse representation into a compact space followed by a sparse point-cloud object detection (SPOD) network to adapt low-density point clouds. Whereas, it causes unaffordable computation overhead in the early-fusion way. Late fusion methods conversely combines independent predictions from diverse vehicles, and conduct proposal refinement to produce the final result [11], [46], [47]. Hurl et al. [11] introduces trust mechanism for secure message selection, and integrates a novel TruPercept to re-weight the output according to consistency score. However, this approach easily suffers from unsatisfactory result due to the over-reliance on individual prediction. For the sake of trade-off between perception accuracy and inference latency, intermediate feature combination among neighboring vehicles has been widely explored

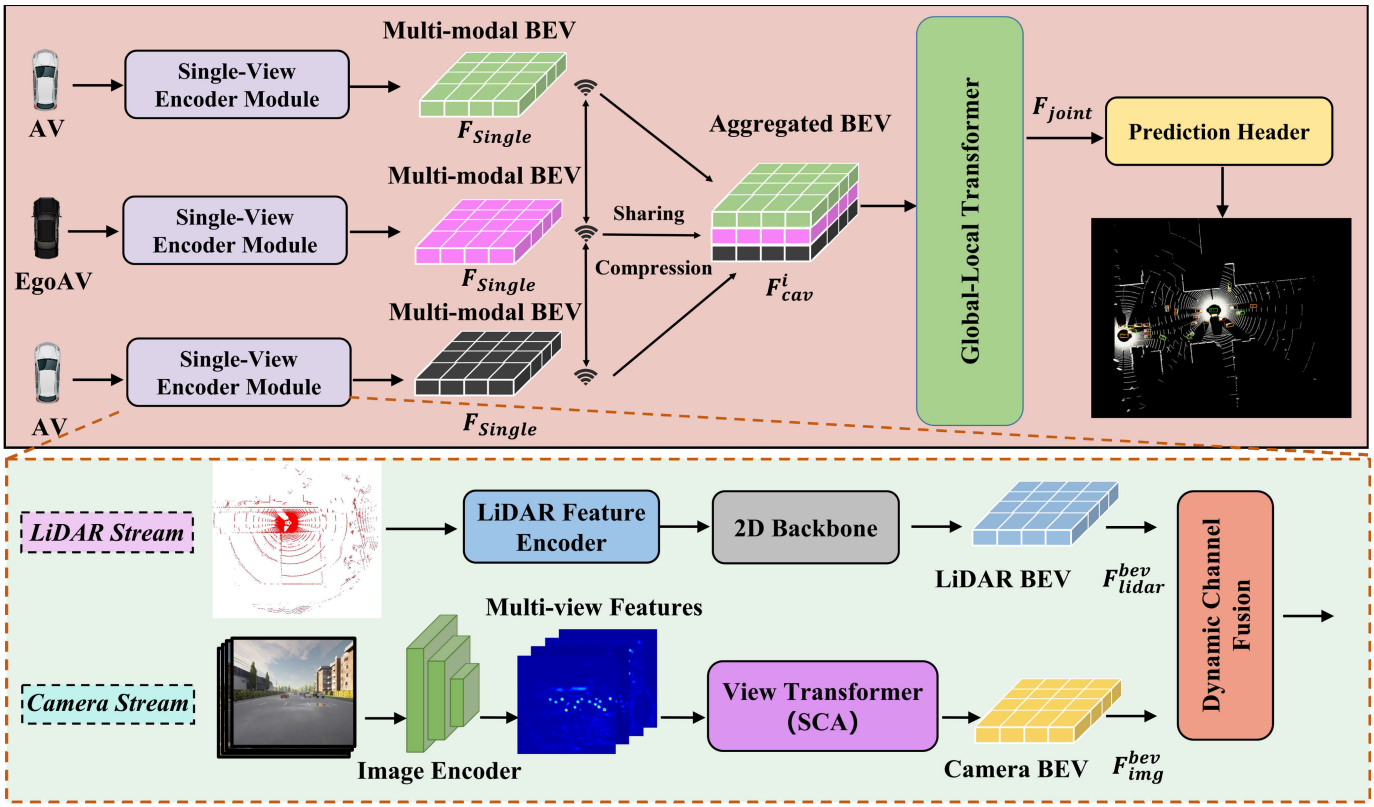


Fig. 2. The architectural diagram of V2VFormer++. For each vehicle, two-stream network with **modality-specific backbone** is adopted for camera-LiDAR feature extraction in BEV plane (with sparse cross-attention **SCA** module for camera-view transformation), and **dynamic channle fusion (DCF)** is designed for fine-grained pixel-point aggregation. Given multi-modal BEV map, data compression and sharing are performed to generate a group of feature map F^i_{cav} at ego-vehicle coordinate. Subsequently, **global-local transformer** collaboration strategy is proposed for channel semantic exploration and spatial correlation modeling among adjacent CAVs. Finally, the multi-vehicle fused map F_{joint} is fed into **prediction header** for object classification and localization regression.

to pursue for favorable performance gains. Wang et al. [14] proposes a graph-based method to iteratively capture and update geographic information for each vehicle by convolutional gated recurrent unit (ConvGRU). To emphasize the agent importance, DiscoNet [15] discards highly-similar pixel among vehicles through an edge weight matrix, and constructs the holistic geometry topology via knowledge distillation. To simulate the effect of transmission latency in the real world, Liu et al. [12] presents a three-step handshake communication protocol including request, match and connect, determining which collaborator to interact with. Moreover, Liu et al. [13] considers a learnable self-attention mechanism to infer whether the ego agent performs an extra communication to obtain more information. Hu et al. [16] develops a novel sparse confidence graph to mask the insignificant element for feature compression. As fine-grained and dense prediction from vehicle-mounted cameras, Xu et al. [48] investigates camera-only map prediction framework under the BEV plane, which utilizes a novel fused axial (FAX) attention to reconstruct dynamic scene on the ground plane. Despite remarkable performance achieved by the abovementioned algorithms, they mostly focus on spatial correlation among CAVs in the local region, without global feature interaction for overlapping semantic refinement. In this work, we attempt to design a novel intermediate feature collaboration dubbed V2VFormer++, that explicitly captures global response among each vehicle, and the ego-networked pair exploits the transformer-based operation for attending to local discriminative feature in a position-wise manner.

III. Methodology

In this section, we would introduce the proposed multi-modal vehicle-to-vehicle cooperative perception framework V2VFormer++. As depicted in Fig. 2, the overall architecture mainly contains four parts: (1) modality-specific backbone for multi-view camera and LiDAR point feature extraction; (2) pixel-point fusion module for semantic and geometrical information aggregation in an adaptative manner; (3) global-local transformer for the informative area attended with self-attention mechanisms; (4) prediction header for producing object localization and classification score.

A. Modality-Specific Backbone

To promote effective feature learning from different modalities, we adopt modality-specific backbone for camera and LiDAR representation extraction. For an individual vehicle, given a set of surrounding-view image as $I = I_1, I_2, \ldots, I_m$, ResNet-style backbone [49] is adopted for comprehensive feature learning from camera image, which contains several strided 3×3 convolutional layers followed by batch normalization (BN) [50] and rectified linear unit (ReLU) [51]. Moreover, the shortcut connection is also constructed with 1×1 convolution, thus enabling stable gradient propagation

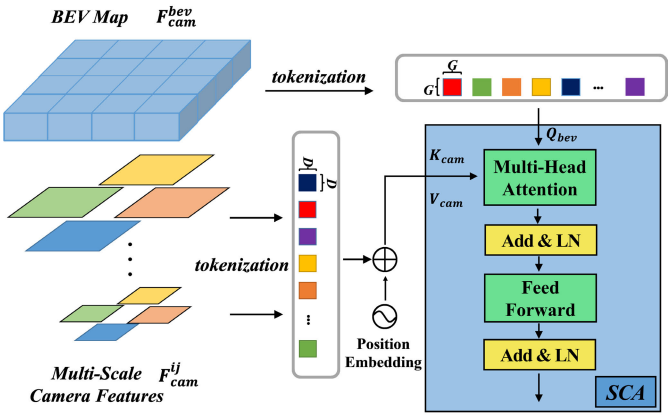


Fig. 3. The schematic diagram of BEV map generation from camera feature. The BEV map F_{cam}^{bev} is initialized by a group of X-Y grids sampled from the world coordinate. Simultaneously, sliding-window sampling method is utilized to partition the BEV and multi-scale camera map F_{cam}^{ij} into a smaller proportion. After linear projection, the query (Q_{bev}), key (K_{cam}) and value (V_{cam}) embeddings are fed into sparse cross-attention (**SCA**) module for iterative BEV map update.

and information delivery. In this way, camera branch produces multi-scale feature map $F_{cam}^{ij} \in \mathbb{R}^{H^j \times W^j \times C_{cam}} (i = 1, \ldots, m; j = 1, \ldots, n)$, where H^j, W^j and C_{cam} denote the height, width and channel number of feature map at different resolutions, and n is the number of feature scale.

Previous works [52], [53], [54] on spatial projection from perspective to bird's-eye view (BEV) spaces explicitly perform depth estimation via camera intrinsic and extrinsic parameters, however, feature ambiguity and inaccurate correspondence inevitably damage the final performance instead. In this work, we primarily sample a group of grids in X-Y plane from the world coordinate, and then project them into the image plane, forming a BEV map F_{cam}^{bev} within the perception range, as shown in Fig. 3. To exploit depth information from various camera setups, a novel sparse cross-attention (SCA) module is adopted for feature interaction between the frontal image and BEV representation. Specifically, an adaptative sliding-window sampling strategy is firstly utilized to partition the resolution of multi-scale feature F_{cam}^{ij} and BEV F_{cam}^{bev} maps into a smaller proportion, with affordable computational overhead. Given the window size $w_1 = D \times D$ and $w_2 = G \times G$ ($G > D$), the resulting feature patch and BEV grid are denoted by $F_{cam}^{ij} \in \mathbb{R}^{\frac{H^j}{D} \times \frac{W^j}{D} \times C_{cam}}$ and $F_{cam}^{bev} \in \mathbb{R}^{\frac{H^{bev}}{G} \times \frac{W^{bev}}{G} \times C_{cam}^{bev}}$, respectively. With independent linear projection, we further produce the query Q_{bev}, key K_{cam} and value V_{cam} from two partitioned sequences, followed by position embeddings to highlight the spatial information. Consequently, sparse cross-attention process can be mathematically described as Eq. 1-Eq. 4:

$$
\begin{aligned}
Q_{bev} &= Linear(F_{cam}^{bev}), \\
K_{cam} &= Linear(F_{cam}^{ij}), \\
V_{cam} &= Linear(F_{cam}^{ij}), \qquad (1) \\
\hat{Q}_{bev} &= SCA(Q_{bev}, K_{cam}, V_{cam}) + Q_{bev} \qquad (2)
\end{aligned}
$$

$$
SCA(Q, K, V) = MultiHead(Concat[\sigma(\frac{Q_t K_t^T}{\sqrt{d}}) V_t])
$$

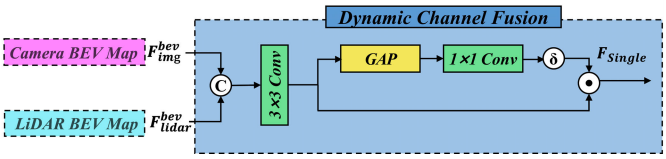


Fig. 4. The schematic diagram of dynamic channel fusion (**DCF**). Given camera-LiDAR BEV maps F_{img}^{bev} and F_{lidar}^{bev}, DCF concatenates them in an element-wise manner, and 3×3 convolution is adopted to explore the valuable semantic and geometric information. After global average pooling operator and MLP (implemented by 1×1 convolution), sigmoid function $\delta(\cdot)$ produces the activation probability for channel feature re-weighting. As a result, multi-modal fused map F_{single} is generated from each single-view encoder module.

$$
t = 1, \cdots, h \qquad (3)
$$

$$
Q_{bev} = \hat{Q}_{bev} + FFN(LN(\hat{Q}_{bev})) \qquad (4)
$$

where $Linear(\cdot)$ is the linear projection with a fully-connected layer, $MultiHead(\cdot)$ is the multi-head self-attention layer, $Concate[\cdot]$ is element-wise feature concatenation, $\sigma(\cdot)$ is the softmax function, h is the head number, $FFN(\cdot)$ defines the feed forward network implemented with multi-layer perceptron, and $LN(\cdot)$ is the layer normalization [55].We conduct three SCA blocks for hierarchical feature aggregation and spatial correlation modeling, and finally, the image BEV map can be expressed as $F_{img}^{bev} \in \mathbb{R}^{H \times W \times C_{img}^{bev}}$.

For the LiDAR branch, we adopt PointPillars [27] backbone for point feature extraction. Denoted the raw point cloud as $P = \{p_1, p_2, \cdots, p_c\}$ ($p_c = (x_c, y_c, z_c, r)$), where x_c, y_c, z_c, r and c represent the spatial coordinates, reflectance and number of point, a stacked pillar tensor is formed with corresponding index, and we utilize a simple PointNet [21] architecture for pillar feature extraction. To generate the pseudo-BEV image, these features are further scattered back to the X-Y plane, and 2D CNN backbone is introduced for merging multi-resolution maps into a dense LiDAR BEV feature $F_{lidar}^{bev} \in \mathbb{R}^{H \times W \times C_{lidar}^{bev}}$.

B. Pixel-Point Fusion Module

Given modal-agnostic BEV representations $F_{img}^{bev} \in \mathbb{R}^{H \times W \times C_{img}^{bev}}$ and $F_{lidar}^{bev} \in \mathbb{R}^{H \times W \times C_{lidar}^{bev}}$, an intuitive idea is to concatenate them together for multi-modal feature enhancement. Nonetheless, it easily suffers from spatial misalignment due to the inherent heterogeneity, and direct concatenation or sum operation generally causes coarse information fusion without fully object semantic supervision. To this end, we design a dynamic channel fusion (DCF) module to exploit image and LiDAR contextual information in a channel-wise manner as depicted in Fig. 4. More specifically, we concatenate each pair of pixel-point feature according to the index, and 3×3 convolution is adopted to explore the valuable semantic and geometric cues, resulting in the re-organized feature F_{conv}. To highlight the object discriminability, global average pooling operator $GAP(\cdot)$ is imposed on feature channel, and a multi-layer perceptron (MLP) with sigmoid function $\delta(\cdot)$ is further utilized to produce the channel activation probability. Finally, we multiply it with the convolved feature F_{conv} to generate the joint feature map $F_{single} \in \mathbb{R}^{H \times W \times C_{lidar}^{bev}}$ with

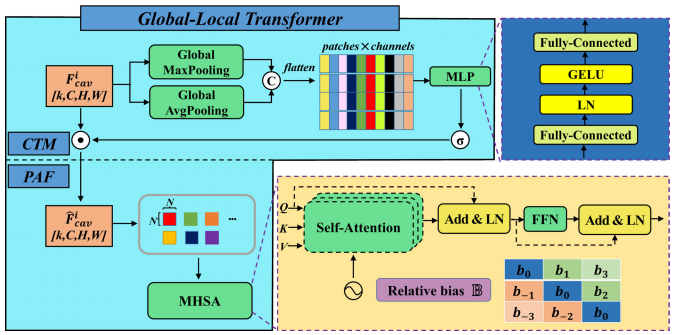


Fig. 5. The schematic diagram of global-local transformer. Given a group of ego-centric feature map F^i_{cav}, channel token mixer (**CTM**) generate a "**patch**×**channels**" table via pooling operators, and after a MLP module, softmax function $\sigma(\cdot)$ outputs the global response value, forming the channel-mixing map $\hat{F}^i_{cav}$. In the position-aware attention fusion (**PAF**), tokenization is performed to partition the embedding into a series of fixed-size window features, and multi-head self-attention (MHSA) is further adopted to explore the spatial correlation of each ego-networked vehicle. Noted that the appended relative bias $\mathbb{B}$ is responsible for contextual relation encoding for each query-key pair.

respect to each vehicle. The whole process can be formulated as Eq. 5:

$$F_{conv} = Conv(Concat[F^{bev}_{img}, F^{bev}_{lidar}])$$
$$F_{single} = \delta(MLP(GAP(F_{conv}))F_{conv} \tag{5}$$

In general, DCF provides an effective solution to exploit channel semantics from both modalities in the unified top-down plane, and this simple module does not damage the inference speed due to its efficient design yet.

C. Global-Local Transformer

For each networked vehicle, we develop an encoder-decoder architecture where the fused map F_{single} is fed into stacked 1×1 convolutions for progressive data compression, and several de-convolutions are accordingly performed for feature recovery, dubbed as $\tilde{F}_{single}$. To compensate spatial-temporal asynchronization, we also adopt an affine transformation $\Phi^i_{cav\to ego}(\tilde{F}_{single})(i = 1, \cdots, k)$ to project different CAV features into the ego-centric view, where $\Phi^i_{cav\to ego}$ denotes the warping function using the sensor calibration matrix, and k is the number of nearby car. As a result, we obtain a group of feature maps $F^i_{cav} = \{F^1_{cav}, \cdots, F_{ego}, F^k_{cav}\}$ at the ego-vehicle coordinate within the communication range, where F_{ego} implies the targeted vehicle.

Previous works typically enhance individual feature map by neither receiving all representations from CAVs in the vicinity nor discarding the entire messages provided by low-relevance collaborators, the former of which unavoidably causes feature redundancy in the overlapping area, while the latter of which probably results in inadequate information interactions across vehicles. To these ends, we propose a novel global-local transformer that consists of channel-token mixer (CTM) for channel semantic filtering and mixing across inter-vehicle patches in a holistic view, and a position-aware attention fusion (PAF) for spatial correlation modeling in the local region. Fig. 5 illustrates the overall structure of global-local transformer.

1) Channel-Token Mixer (CTM): The two-stage transformation can be referred to tokenization and mixing process. Given the CAV feature maps $F^i_{cav} = \{F^1_{cav}, \cdots, F_{ego}, F^k_{cav}\} \in \mathbb{R}^{H\times W\times C^{bev}_{lidar}\times k}$, we primarily leverage 3D feature pooling operator (i.e., global max-pooling $(GMP(\cdot))$ and global average pooling $(GAP(\cdot))$ to reflect the channel-wise information particularity and commonality, respectively. Then, feature vectorization is performed by concatenating and flattening $(flatten(\cdot))$ them into a sequence of image tokens, forming a "**patches**×**channels**" table $T \in \mathbb{R}^{S\times C}(S = 1 \times 1 \times 1, C = 2k)$. The whole process can be described as Eq.6:

$$T = Flatten(Concat[GMP(F^i_{cav}), GAP(F^i_{cav})]) \tag{6}$$

Subsequently, Mixer conducts linear feature projection into a hidden space via a two-layer MLP, followed by layer normalization and Gaussian error linear unit $(GELU(\cdot))$. It acts on the rows of table T, maps $\mathbb{R}^C \to \mathbb{R}^{C_{hid}} \to \mathbb{R}^C$, and shares information across all rows, thus promoting for channel communication. Finally, sofxmax function is applied for channel-wise importance evaluation, and we multiply it with the CAV maps in an element-wise manner, as formulated in Eq.7:

$$\hat{F}^i_{cav} = \sigma(W_2 \bigotimes GELU(LN((W_1 \bigotimes T)_{*,j})))F^i_{cav},$$
$$i = 1, \cdots, k$$
$$j = 1, \cdots, C_{hid} \tag{7}$$

where W_* denotes the weight of linear projection, $\bigotimes$ is the matrix mulplication, $()_{*,j}$ implies the operator on feature channel, C_{hid} is the tunable channel number in the hidden layer and $\hat{F}^i_{cav} \in \mathbb{R}^{H\times W\times C^{bev}_{lidar}\times k}$ denotes the channel-mixing feature map. Benefited by the powerful of MLP, CTM is capable of dynamically filtering the irrelevant representation (i.e., overlapping signal), meanwhile capturing the global response scattered in per-location map. More importantly, it strategically performs feature mixing across channel to enhance the valuable information expression, with significant memory savings.

2) Position-Aware Attention Fusion (PAF): To further capture the long-range dependencies among vehicles, transformer-based architecture is widely applied with self-attention mechanism to explore spatial relationship of each ego-networked map. Nonetheless, it requires much longer training epochs for convergence, and dense dot-product operation brings unbearable computational budget. In this work, we design a position-aware attention fusion (PAF) module, that is composed of sparse window-based tokenization and self-attention mechanism with relative offset for local feature interactions across all locations. Formally, channel-mixing map $\hat{F}^i_{cav} \in \mathbb{R}^{H\times W\times C^{bev}_{lidar}\times k}$ is linearly projected into the high-dimensional space to generate three feature embeddings $F^e \in \mathbb{R}^{H\times W\times C}$ $(e = 1, 2, 3$ and $C = C^{bev}_{lidar}$ for brevity). Subsequently, we partition them into a series of 3D non-overlapping window w_3 with a size of $N \times N$, respectively, forming F^1_{win}, F^2_{win} and F^3_{win} are in the same dimension $(\frac{H}{N} \times \frac{W}{N}) \times (N \times N \times k) \times C$. It is highlighted that window-level partition can reach an effective tokenization than

intensive computation on the per-pixel map. Consequently, each token is flattened to generate a sequence of query (Q), key (K) and value (V), and we further introduce a multi-head self-attention ($MHSA(\cdot)$) layer with relative bias $\mathbb{B}$ to explore intra- and inter-vehicle spatial correlation. Analogous to position embedding (PE), $\mathbb{B}$ is a fixed-size window index responsible for contextual relation learning from each query-key pair. Mathematically, PAF procedure can be described as Eq. 8-Eq. 11:

$$\begin{aligned} F^e_{win} &= Window[Linear(\hat{F}^i_{cav})], e = 1,2,3 \\ Q &= Flatten(F^1_{win}), \\ K &= Flatten(F^2_{win}), \\ V &= Flatten(F^3_{win}), \end{aligned} \tag{8}$$

$$\hat{Q} = MHSA(Q,K,V) + Q \tag{9}$$

$$\begin{aligned} MHSA(Q,K,V) = MultiHead \\ (Concate[\sigma(\frac{Q_t K_t^T}{\sqrt{d}} + \mathbb{B})V_t]) \\ t = 1, \cdots, h \end{aligned} \tag{10}$$

$$Q = \hat{Q} + FFN(LN(\hat{Q})) \tag{11}$$

where $Window[\cdot]$ implies the window-level patch partition. We utilize two-layer self-attention operation for exploiting fine-grained position information, and multi-vehicle fused map can be referred to as $F_{joint} \in \mathbb{R}^{H \times W \times C \times k}$. Taken the advantage of window-level attention, PAF module is not only robust to pose estimation and offset error, but the contour-aware attribute (e.g., edge and boundary) can also facilitate the detection performance on hard object.

D. Prediction Header

As commonly done [14], [16], [17], [48], the joint feature map F_{joint} is then fed into classification and regression heads for object category and localization prediction, respectively. Notably, the post process with non-maximum suppression (NMS) is adopted for redundant proposal removal.

During model training, the loss function $\mathcal{L}$ contains classification $\mathcal{L}_{cls}$ and regression $\mathcal{L}_{reg}$ parts. Given the ground-truth box $B_{gt} = (x, y, z, w, l, h, \theta)$, where (x, y, z) denotes the object center, (w, l, h) defines the 3D box dimension, and θ is the heading orientation, we adopt focal loss [56] ($FL(\cdot)$) to balance the background-foreground sample, and smooth-$L1$ function is utilized for supervising 3D box size. Detailed information can refer to Eq. 12-Eq. 14:

$$\mathcal{L} = \beta_1 L_{cls} + \beta_2 L_{reg} \tag{12}$$

$$FL(p_t) = -\alpha_t (1-p_t)^{\gamma} \log(p_t) \tag{13}$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & if|x| < 1, \\ |x| - 0.5, & otherwise. \end{cases} \tag{14}$$

where $\beta_1 = 1.0$ and $\beta_2 = 2.0$ are the weight parameters, α and γ are the hyperparameter of focal loss, p_t is the estimated softmax probability. Noted that heading orientation θ is encoded by sinusoidal function (i.e., $sin(\theta_{gt} - \theta_{pred})$) before smooth-$L1$ computation, where θ_{gt} and θ_{pred} present the ground-truth and predicted angles, respectively.

IV. Experiments

In this section, quantitative and qualitative experiments are conducted on vehicle-to-vehicle perception benchmarks to investigate the effectiveness of our proposed framework and its components. Detailed information (i.e, dataset, implementation, ablation study, etc.) would be described as follows.

A. Datasets

OPV2V is a large-scale vehicle-to-vehicle collaborative perception dataset, which is built on the top of OpenCDA platform [57] and CARLA simulator [58]. Generally, it contains $12k$ frames of 3D point clouds and RGB images generated by four vehicle-mounted cameras and one 64-channels LiDAR sensor, with $230k$ 3D box annotations covered a full 360° view. In our experiment, the detection range is set to [-64,64] m, [-40,40] m and [-3,1] m along the x, y and z axes, respectively. The model is trained and validated with 6765 and 1980 samples, and we test the final cooperative performance on 2170 *Default* and 550 *Culver City* splits.

V2X-Sim 2.0 is a synthesized multi-modal benchmark for vehicle-to-everything (V2X) perception evaluation, co-simulated by CARLA and micro-traffic simulator SUMO [59]. It is composed of 100 scenes durated a 20-second traffic flow at the intersection of three CARLA towns, with $37.2k$ training, $5k$ validation and $5k$ test data. Each scene has 2-5 CAVs equipped with six cameras and one 32-channel LiDAR, as well as GPU and IMU sensors. Similarly, the perception area is limited to [-32,32]$m\times$[-32,32]$m\times$ [-3,2]m in our study.

B. Implementation Details

The experimental platform is based on 8 NVIDIA Tesla V100 GPUs, and we define the communication range as $70m$ by default. For OPV2V [19], we introduce the curriculum learning strategy [60] to imitate the human cognition mechanism: the model is trained for 35 epochs at *sim* mode and another 10 epochs with *real* setting (e.g., localization error, async overhead, etc.), optimized by Adam [61] with 0.0002 initial learning rate, 0.02 weight decay and cosine learning rate scheduler. Besides, several tricks (i.e., warmup and early-stop) are also adopted for training stability, and the score and IoU thresholds for NMS post-processing are set to 0.6 and 0.15, respectively. As for V2X-Sim 2.0 [20], we follow DiscoNet settings, and technique details can refer to [15]. The score and IoU thresholds for NMS procedure are set to 0.6 and 0.15.

The image cropped with a reslution of 520×520 pixels is fed into ResNet-34 [49] encoder for multi-scale feature extraction, and the generated BEV grid is $0.25m$. We take four attention heads ($h = 4$), and the window size $D = (8, 8, 16)$ and $G = (16, 16, 32)$ in hierarchical SCA module. Moreover, the voxel size is set to $(0.25, 0.25, 4)$ along x-y-z axis, and in global-local transformer, the window size N is 4. Unless otherwise stated, we report the 3D detection average precision (AP) at 0.5 and 0.7 IoU thresholds for a fair comparison.

TABLE I

DETECTION RESULTS ACHIEVED BY COBEVT [48], WHERE2COMM [16], V2VNET [14] AND V2VFORMER++ ON OPV2V TEST SPLITS, WE HIGHLIGHT THE BEST ACCURACY AT 0.5 AND 0.7 IOU THRESHOLDS WITH BOLD FONT

Methods	Modality	*Default*		*Culver City*	
		AP@0.5(%)	AP@0.7(%)	AP@0.5(%)	AP@0.7(%)
CoBEVT [48]	L	88.3	79.7	80.4	67.2
Where2comm [16]	L	88.9	74.5	**82.7**	68.0
V2VFormer++-L(ours)	L	**88.9**	**82.0**	80.5	**70.0**
V2VNet [14]	L+C	92.6	87.8	89.7	83.0
CoBEVT [48]	L+C	92.9	89.0	90.6	84.2
Where2comm [16]	L+C	93.2	89.4	**92.5**	**85.4**
V2VFormer++(ours)	L+C	**93.5**	**89.5**	92.3	85.2

TABLE II

DETECTION RESULTS ACHIEVED BY WHEN2COM [13], WHO2COM [12], V2VNET [14], DISCONET [15] AND V2VFORMER++ ON V2X-SIM 2.0 TEST SET. BESIDES, WE LIST THE UPPER-BOUND AND LOWER-BOUND PERFORMANCE, AND THE BEST ACCURACY AT 0.5 AND 0.7 IOU THRESHOLDS IS ALSO HIGHLIGHTED WITH BOLD FONT

Methods	Modality	Collaboration Approach			Average Precision(AP)	
		Early	Intermediate	Late	IoU=0.5	IoU=0.7
Upper-bound	L	✓	✗	✗	63.3	60.2
		✓	✗	✓	59.7	55.8
When2com [13]	L	✗	✓	✗	45.7	41.8
Who2com [12]	L	✗	✓	✗	44.8	40.4
V2VNet [14]	L	✗	✓	✗	56.8	50.7
DiscoNet [15]	L	✗	✓	✗	60.3	53.9
Lower-bound	L	✗	✗	✓	57.6	54.2
		✗	✗	✗	45.8	42.3
V2VFormer++(ours)	L+C	✗	✓	✗	**72.7**	**65.5**

C. Quantitative Results

Table I illustrates the cooperative perception result of our proposed V2VFormer++ and four counterparts on both OPV2V *Default* and *Culver City* splits. On the one hand, we remove the camera stream from each single-view module, and evaluate LiDAR-only detection performance, dubbed as V2VFormer++-L. It is observed that our proposed method outperforms the CoBEVT [48] and Where2comm [16] methods over considerable performance gains of 2.3% $\sim$ 7.5% and 2.0% $\sim$ 2.8% AP@0.7 at *Default* and *Culver City* sets, suggesting its effectiveness and superiority. On the other hand, we append the same camera stream into the LiDAR-only cooperative detectors (i.e., V2VNet [14], CoBEVT [48], Where2comm [16]), and assess the multi-modal detection accuracy. Our proposed V2VFormer++ reports the top collaborative perception performance: it achieves 93.5% and 89.5% AP at 0.5 and 0.7 IoU thresholds on *Default*, which surpasses three alternatives by 0.3% $\sim$ 0.9% AP@0.5 and 0.1% $\sim$ 1.7% AP@0.7. Moreover, V2VFormer++ is on par with the first-tier Where2comm [16] (only 0.2% AP behind) on *Culver City*, demonstrating its competitiveness and adaptation.

Simultaneously, collaborative detection result on V2X-Sim 2.0 test set is also tabulated in Table II, and we reproduce different fusion strategies (e.g., Early, Intermediate and Late) based on the DiscoNet [15]. Without bells and whistles, our V2VFormer++ achieves the state-of-the-art cooperative detection accuracy with 72.7% AP@0.5 and 65.5 AP@0.7, respectively. Compared with the other intermediate counterparts (e.g., DiscoNet [15]), V2VFormer++ receives more than 10% AP boosts at both two IoU thresholds, implying the advancement of proposed feature collaboration. Furthermore, it outperforms the Upper-bound by 9.4% AP@0.5 and 5.3% AP@0.7 margins. We argued that the model fails to leverage the meaningful information from adjacent CAVs due to noisy raw point cloud, while our intermediate representation provides rich object semantic and geometric information for faciliating collaborative perception performance reasonably.

D. Ablation Study

For simplicity, ablation study would be investigated on OPV2V *Default* and *Culver City* splits, to measure the effectiveness and robustness of our proposed framwork.

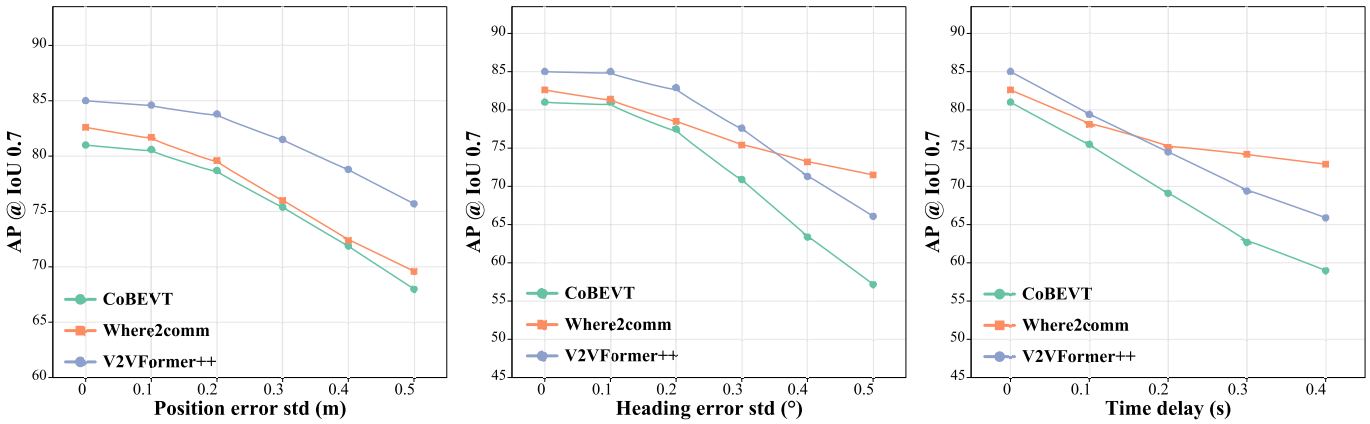


Fig. 6. Ablation study of the robustness test. Notably, all experiments are conducted on OPV2V *Default* split. **Left.** The relationship between position error and AP result at 0.7 IoU threshold; **Middle.** The relationship between heading error and AP result at 0.7 IoU threshold; **Right.** The relationship between time delay and AP result at 0.7 IoU threshold.

TABLE III

ABLATION STUDY OF THE EFFECTIVENESS OF INDIVIDUAL COMPONENTS ON OPV2V TEST SPLITS. THE ACCURACY GAINS/ DROPS ARE HIGHLIGHTED WITH DIFFERENT COLORS IN THE BRACKETS, RESPECTIVELY

DCF	Global-Local Transformer	Default		Culver City	
		AP@0.5(%)	AP@0.7(%)	AP@0.5(%)	AP@0.7(%)
		85.0	72.0	80.9	64.0
✓		92.6(**+7.6**)	87.8 (**+15.8**)	89.7(**+8.8**)	83.0(**+19.0**)
	✓	88.9(**+3.9**)	82.0(**+10**)	80.5	70.0(**+6.0**)
✓	✓	93.5(**+8.5**)	89.5(**+17.5**)	92.3(**+11.4**)	85.2(**+21.2**)

TABLE IV

ABLATION STUDY OF PERFORMACE CONTRIBUTION PROBIDED BY MULTI-MODAL FUSION METHODS ON OPV2V TEST SPLITS. THE ACCURACY GAINS/DROPS ARE HIGHLIGHTED WITH DIFFERENT COLORS IN THE BRACKETS, RESPECTIVELY

	Methods	*Default*		*Culver City*	
		AP@0.5(%)	AP@0.7(%)	AP@0.5(%)	AP@0.7(%)
Concat	V2VNet [14]	92.0	87.0	90.0	79.0
	CoBEVT [48]	92.9	88.3	91.0	84.0
	Where2comm [16]	92.2	88.1	92.1	83.8
	V2VFormer++(ours)	93.2	89.1	91.2	85.1
DCF	V2VNet [14]	92.6(**+0.6**)	87.8(**+0.8**)	89.7(**-0.3**)	83.0(**+4.0**)
	CoBEVT [48]	92.9	89.0(**+0.7**)	90.6(**-0.4**)	84.2(**+0.4**)
	Where2comm [16]	93.2(**+1.0**)	89.4(**+1.3**)	92.5(**+0.4**)	85.4(**+1.6**)
	V2VFormer++(ours)	93.6(**+0.4**)	89.5(**+0.4**)	92.3(**+1.1**)	85.2(**+0.2**)

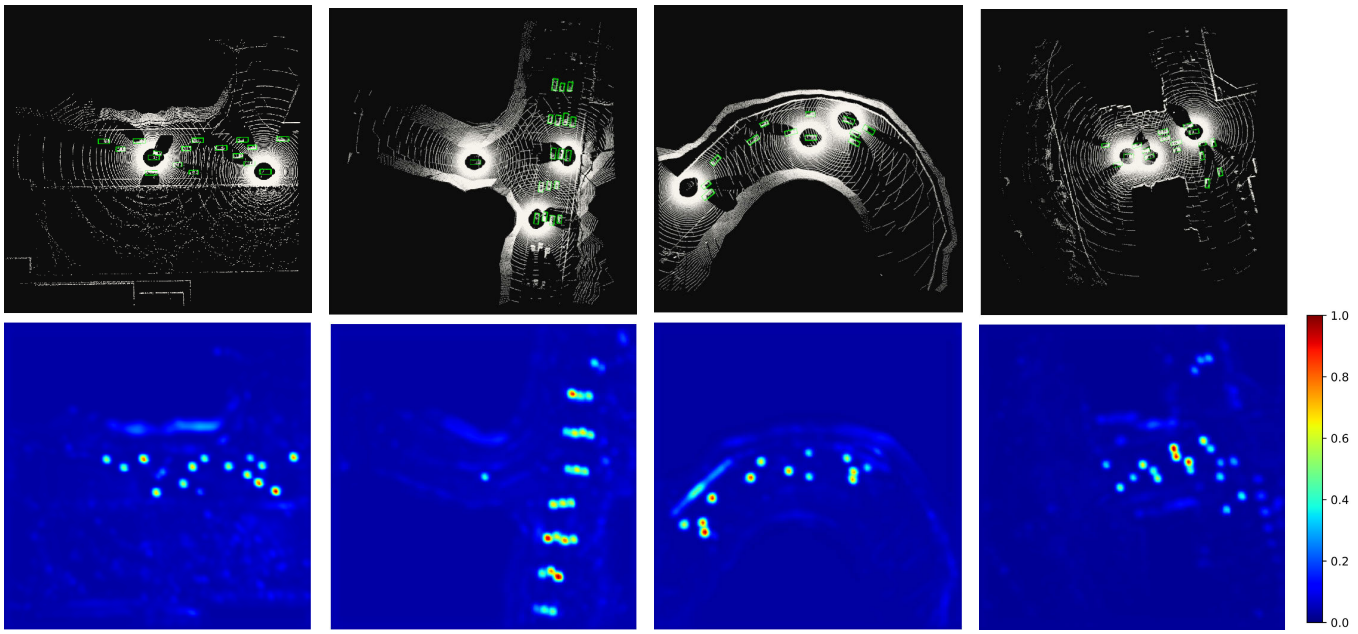
1) Effectiveness of Component: For clarify, we choose V2VNet [14] as the baseline that acheives 85.0% AP@0.5 and 72.0% AP@0.7% in *Default*, 80.9% AP@0.5 and 64.0% AP@0.7 in *Culver City*, respectively, as tabulaetd in Table III. When appending camera branch with dynamic channel fusion (DCF), it provides 7.6% $\sim$ 19.0% accuracy gain at 0.5 and 0.7 IoU thresholds. Furthermore, we replace the spatially aware graph neural network(GNN) proposed by V2VNet [14] with global-local transformer to measure its contribution to collaborative perception. Likewise, it offers 10.0% AP@0.7 gains in *Default*. Finally, V2VFormer++ incorporates DCF with global-local transformer into the baseline, and the best performance is observed, demonstrating the effectiveness of each component.

To further investigate heterogeneous data fusion under the single-vehicle view, we extend the LiDAR-only detectors (e.g., V2VNet [14], CoBEVT [48] and Where2comm [16]) with camera stream, and adopt two camera-LiDAR aggregation methods for comparison. As shown in Table IV, dynamic channel fusion (**DCF**) presents a better multi-modal feature combination than concatenation (**Concate**) among different collaboration frameworks: despite slight performance drop, it steadily provides 0.4% $\sim$ 1.0% AP@0.5 increments and 0.4% $\sim$ 4.0% AP@0.7 promotions on both *Default* and *Culver City* sets. Benefited by channel pooling and re-weighting operations, DCF is able to fully exploit semantic and geometric information from various modalities, and expressive multi-modal representation is favorable for cooperative performance enhancement.

2) Robustness Test: To analyze the robustness of cooperative perception, we firstly adopt curriculum learning on several multi-modal perception frameworks, and list the cooperative result of OPV2V *Default* set under different mode configurations as tabulated in Table V. Notably, **Sim/Real** modes define the desirable/real-world transmission without/with data compression, while **Perfect/Noisy** conditions stand for the ideal/corrupted environments without/with localization error (e.g., Gaussion noise) and communication latency (uniform distribution over 200ms), respectively.

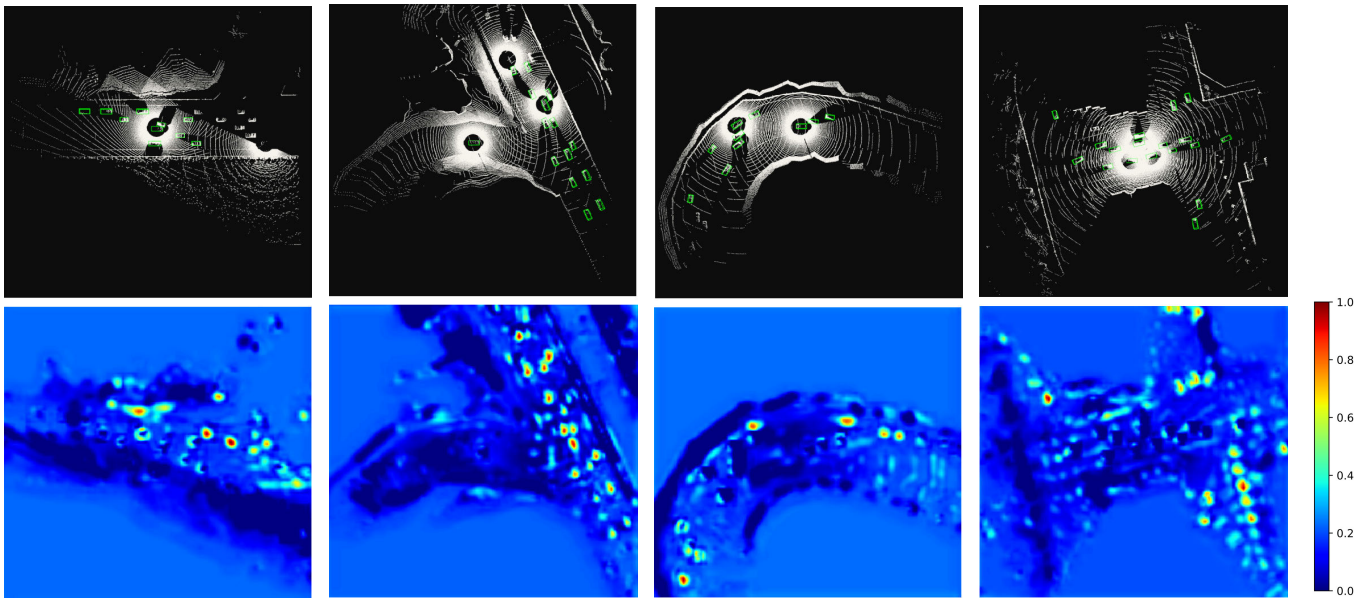
Obviously, the proposed V2VFormer++ shows the strong robustness against different corruptions: it achieves a favorable detection accuracy of 84.9% AP@0.5 and 58.5% AP@0.7 on the **Sim+Noisy** level, and offers 6.0% and 16.9% AP gains under the **Real** environment. When transferring from **Perfect** to **Noisy** in the **Real** setting, all collaborations suffer from a subtantial accuracy decline, e.g., 6.3% AP@0.5 and 10.9% AP0.7 in V2VNet. Our V2VFormer++ reports an acceptable performance drops in 2.1% AP@0.5 and 9.6% AP@0.7, suggesting the perferable stability and generality.

We further add Gaussian noise and uniform distribution to simulate different real disturbances, and the anti-interference ability of position error, heading error and communication delay is verified as illustrated in Fig. 6. Evidently, our proposed method reveals the remarkable and advantageous performance against localization offset over Gaussion distribution with standard deviation (std) $\sigma_{xyz} \in [0, 0.5]m$, while the counterparts (e.g., CoBEVT [48]) experiences an appearant performance



(a) Straight (b) Merge (c) Curve (d) Intersection

Fig. 7. Visualization results of attention map activated by dynamic channel fusion (DCF) module. Four common scenarios (i.e., **Straight**, **Merge**, **Curve** and **Intersection**) are randomly selected from the left to the right, and a pair of LiDAR ground-truth (GT) and attention map is correspondingly listed in each column. Noted that a spot with larger activated value implies higher potential of target ocurred in this area.



(a) Straight (b) Merge (c) Curve (d) Intersection

Fig. 8. Visualization results of attention map activated by global-local transformer collaboration strategy. Four common scenarios (i.e., **Straight**, **Merge**, **Curve** and **Intersection**) are randomly selected from the left to the right, and a pair of LiDAR ground-truth (GT) and attention map is correspondingly listed in each column. Noted that a spot with larger activated value implies higher potential of target ocurred in this area.

decline with the increasing offset value. Besides, it is non-susceptible to varying heading noise with std $\sigma_{ryp} \in [0°, 1°]$, and also maintains favorable AP results under $[0, 400]ms$ time delay. Generally, it is suggested that V2VFormer++ holds the outstanding robustness and anti-interference ability confronted with severe real scenarios. Owing to curriculum learning strategy, the model could explore inherent and significant information step-by-step, and we deem these knowledges would faciliate to retain considerable perception performance even under various corruptions. More importantly, global-local transformer collaboration strategy incorporates the ego-centric perspective with multi-view representations organically, which is helpful to hard-sample perception in such the occluded and line-of-beyond areas.

E. Qualitative Results

Finally, qualitative experiments are performed to deeply analyze how the fuison module make effect. And we also

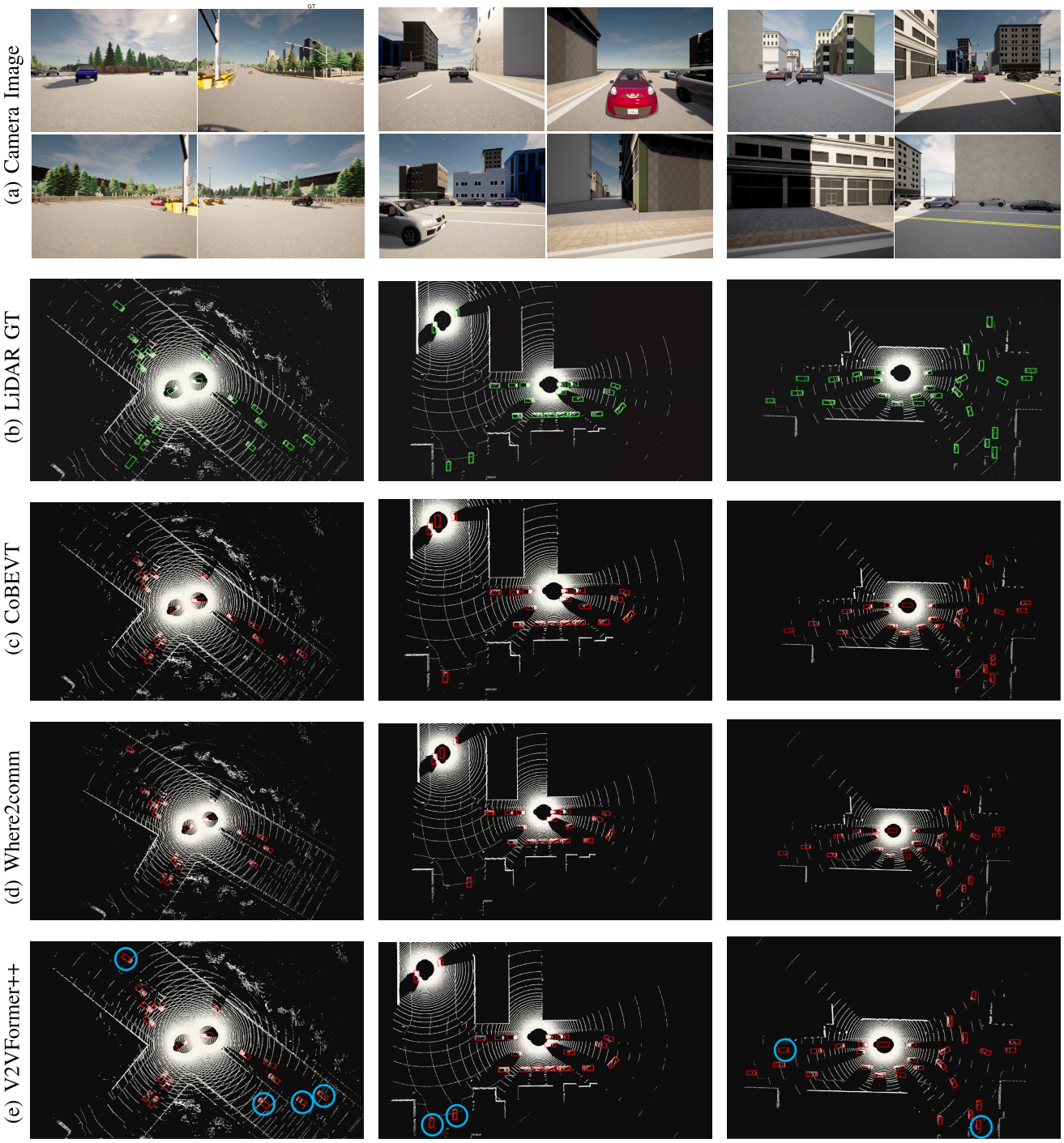


Fig. 9. Detection visualizations in OPV2V test split. For top to bottom, we list the original camera image, LiDAR ground-truth (GT), and perception results achieved by CoBEVT [48], Where2comm [16] and V2VFormer++. Noted that the GT and predicted boxes are drawn with red and green colors, respectively, and we also highlight the superiority and advancement of V2VFormer++ with blue circle. Evidently, our proposed method shows more accurate and robust collaborative detection performance compared to other methods, even in heavily occluded, blind-spot and line-of-beyond areas.

showcase the cooperative detection results to reflect the advantage of our proposed V2VFormer++.

1) Attention Map: As depicted in Fig. 7 and Fig. 8, we seperately exhibit a pair of LiDAR ground-truth (GT) and activation map after DCF and global-local transformer collaboration at the straight, merge, curve, intersection situations. Thanks to the effective DCF design, the activated spot can approximately correspond to the target area in the LiDAR GT, allowing the model to focus on the high potential or region of interest (RoI) of object. DCF explores feature channel semantics in a dynamic point-wise manner, and thus valuable information could be fully expoited from camera and LiDAR modalities. Similarly, global-local transformer consumes multi-view representations from adjacent CAVs via channel-wise and position-aware importances. It would provide a broader and longer probing range, and the highlighted spots could guide the model to detect the occluded or rarely-seen objects.

TABLE V

ABLATION STUDY OF MODEL ROBUSTNESS AGAINST VARIOUS MODE CONFIGURATIONS ON OPV2V *Default* SPLIT. THE ACCURACY GAINS ARE HIGHLIGHTED IN THE BRACKETS, RESPECTIVELY

	Methods	Perfect		Noisy	
		AP@0.5(%)	AP@0.7(%)	AP@0.5(%)	AP@0.7(%)
Sim	V2VNet [14]	-	-	77.4	53.9
	CoBEVT [48]	-	-	83.2	51.0
	Where2comm [16]	-	-	86.0	60.0
	V2VFormer++(ours)	-	-	84.9	58.5
Real	V2VNet [14]	90.8	81.9	84.5(**+7.1**)	71.0(**+17.1**)
	CoBEVT [48]	90.0	81.2	87.9(**+4.7**)	70.3(**+19.3**)
	Where2comm [16]	90.4	82.0	89.1(**+3.1**)	75.5(**+15.5**)
	V2VFormer++(ours)	93.0	85.0	90.9(**+6.0**)	75.4(**+16.9**)

2) Detection Visualization: As listed in Fig. 9, we display the visualizations in comparison of our V2VFormer++ with CoBEVT [48] and Where2comm [16] methods. Typically, our proposed algorithm consistently maintains much precise and robust detection results, particularly in challenging and ambinguous scenes. It still shows outstanding perception ability in those hard samples (i.e., occluded, blind-spot and line-of-beyond areas) that other counterparts fails, suggesting its supriority and advancement.

V. CONCLUSION

In this paper, we make the first attempt to vehicle-to-vehicle cooperative framework with multi-modal representation, dubbed as V2VFormer++. For individual vehicle, two-stream architecture with sparse cross-attention (SCA) transformation and dynamic channel fusion (DCF) is proposed for camera-LiDAR feature aggregation under the unified bird's-eye-view (BEV) space, thus exploiting semantic and geometric informaiton fully. To leverage inter-vehicle correlation from adjacent CAVs better, we design a two-stage global-local transformer collaboration strategy where channel token mixer (CTM) captures the global response scattered in per-location map and postion-aware fusion (PAF) explores the spatial relationship of each ego-networked pair in the local perspective. Empirical experiments are conducted on both OPV2V [19] and V2X-Sim 2.0 [20] benchmarks, and the results demonstrate our proposed V2VFormer++ ourperforms all counterparts by a substantial margin, suggesting its effectiveness and superiority. Moreover, ablation study and visualization analysis further reveal the strong robustness against various disturbances from real-world scenarios.

Future works would cotinue to investigate how adverse factors influent make impact on multi-agent percepton algorithm, e.g., latency, lossy package, etc. Also, how to optimize inference efficiency is comparatively essential for the practical deployment.

REFERENCES

[1] P. Sun and A. Boukerche, "Toward the design of an efficient transparent traffic environment based on vehicular edge computing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[2] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[3] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query DeNoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13609–13617.

[4] P. Adarsh, P. Rathi, and M. Kumar, "YOLO v3-tiny: Object detection and recognition using one stage improved model," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 687–694.

[5] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "CL3D: Camera-LiDAR 3D object detection with point feature enhancement and point-guided fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 18040–18050, Oct. 2022.

[6] A. Ganeshan et al., "Warp-refine propagation: Semi-supervised auto-labeling via cycle-consistency," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15479–15489.

[7] H. Perreault, G.-A. Bilodeau, N. Saunier, and M. Héritier, "CenterPoly: Real-time instance segmentation using bounding polygons," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2982–2991.

[8] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.

[9] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2019, pp. 514–524.

[10] 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 *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 88–100.

[11] B. Hurl, R. Cohen, K. Czarnecki, and S. Waslander, "TruPercept: Trust modelling for autonomous vehicle cooperative perception from synthetic data," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Oct. 2020, pp. 341–347.

[12] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6876–6883.

[13] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4105–4114.

[14] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 605–621.

[15] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 29541–29552.

[16] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. 36th Conf. Neural Inf. Process. Syst. (Neurips)*, Nov. 2022, pp. 4874–4886.

[17] R. Xu, X. Xiang, Z. Tu, X. Xia, M. H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2022, pp. 107–124.

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

[19] 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 *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589.

[20] Y. Li et al., "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.

[21] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.

[22] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5099–5108.

[23] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3DSSD: Point-based 3D single stage object detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11037–11045.

[24] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, and Y. Guo, "Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18931–18940.

[25] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[26] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[27] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12689–12697.

[28] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[29] J. Deng, S. Shi, D. Xiang, X. Chen, and H. Huang, "Voxel R-CNN: Towards high performance voxel-based 3D object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 35, 2021, pp. 955–963.

[30] H. Wang et al., "CAGroup3D: Class-aware grouping for 3D object detection on point clouds," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 29975–29988.

[31] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "STD: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1951–1960.

[32] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10526–10535.

[33] S. Shi et al., "PV-RCNN++: Point-Voxel feature set abstraction with local vector representation for 3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 2, pp. 531–551, Feb. 2023.

[34] S. Shi, X. Wang, and H. Li, "Part-A^2 Net: 3D part-aware and aggregation neural network for object detection from point cloud," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 4, pp. 765–781, Mar. 2020.

[35] W. Zheng, W. Tang, L. Jiang, and C.-W. Fu, "SE-SSD: Self-ensembling single-stage object detector from point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14489–14498.

[36] J. Noh, S. Lee, and B. Ham, "HVPR: Hybrid voxel-point representation for single-stage 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14600–14609.

[37] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4603–4611.

[38] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3D-CVF: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3D object detection," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 720–736.

[39] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3D object detection," in *Proc. ECCV*, Oct. 2022, pp. 628–644.

[40] Z. Yang, J. Chen, Z. Miao, W. Li, X. Zhu, and L. Zhang, "DeepInteraction: 3D object detection via modality interaction," in *Proc. NeurIPS*, 2022, pp. 1992–2005.

[41] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3D detection," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 16494–16507.

[42] Y. Jiao, Z. Jie, S. Chen, J. Chen, L. Ma, and Y.-G. Jiang, "MSMDFusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21643–21652.

[43] X. Bai et al., "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1080–1089.

[44] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3D object detection," 2022, *arXiv:2206.00630*.

[45] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," 2022, *arXiv:2205.13542*.

[46] Z. Y. Rawashdeh and Z. Wang, "Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3961–3966.

[47] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 953–959.

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

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[51] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[52] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[53] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.

[54] H. Zhou, Z. Ge, Z. Li, and X. Zhang, "MatrixVT: Efficient multi-camera to BEV transformation for 3D perception," 2022, *arXiv:2211.10593*.

[55] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[56] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[57] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "OpenCDA: An open cooperative driving automation framework integrated with co-simulation," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 1155–1162.

[58] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Annu. Conf. Robot Learn.*, 2017, pp. 1–16.

[59] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo-simulation of urban mobility: An overview," in *Proc. Int. Conf. Adv. Syst. Simulation*, 2011, pp. 23–28.

[60] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

Hongbo Yin received the B.E. degree in traffic engineering from Southwest Jiaotong University in 2022. He is currently pursuing the master's degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include autonomous driving, computer vision, cooperative perception, and intelligent transportation systems.

Daxin Tian (Senior Member, IEEE) is currently a Professor with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include mobile computing, intelligent transportation systems, vehicular ad hoc networks, and swarm intelligence. He is an IEEE Intelligent Transportation Systems Society Member and an IEEE Vehicular Technology Society Member.

Chunmian Lin received the Ph.D. degree in electronics and information from Beihang University, Beijing, China. He is currently a Post-Doctoral Researcher with the School of Transportation Science and Engineering, Beihang University. His current research interests include autonomous driving, image processing, computer vision, artificial intelligence, and deep learning, particularly their applications in intelligent transportation systems.

Xuting Duan (Member, IEEE) received the Ph.D. degree in traffic information engineering and control from Beihang University, Beijing, China. He is currently an Assistant Professor with the School of Transportation Science and Engineering, Beihang University. His current research interests include vehicular ad hoc networks, cooperative vehicle infrastructure systems, and the Internet of Vehicles.

Jianshan Zhou received the B.Sc. and M.Sc. degrees in traffic information engineering and control from Beihang University, Beijing, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree with the School of Transportation Science and Engineering. His current research interests include wireless communication, artificial intelligent systems, and intelligent transportation systems.

Dezong Zhao (Senior Member, IEEE) received the B.Eng. and M.S. degrees from Shandong University, Jinan, China, in 2003 and 2006, respectively, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2010, all in control science and engineering. He is a Senior Lecturer in autonomous systems with the School of Engineering, University of Glasgow, U.K. His research interests include connected and autonomous vehicles, machine learning, and control engineering. His work has been recognized by being awarded an EPSRC Innovation Fellowship and a Royal Society-Newton Advanced Fellowship in 2018 and 2020, respectively.

Dongpu Cao received the Ph.D. degree from Concordia University, Canada, in 2008. He is the Canada Research Chair in driver cognition and automated driving and an Associate Professor and the Director of Waterloo Cognitive Autonomous Driving (CogDrive) Laboratory, University of Waterloo, Canada. He has contributed more than 200 papers and three books. His current research interests include driver cognition, automated driving, and cognitive autonomous driving. He received the SAE Arch T. Colwell Merit Award in 2012, the IEEE VTS 2020 Best Vehicular Electronics Paper Award, and three Best Paper Awards from the ASME and IEEE conferences. He is an IEEE VTS Distinguished Lecturer. He serves on the SAE Vehicle Dynamics Standards Committee and acts as the Co-Chair of the IEEE ITSS Technical Committee on Cooperative Driving. He serves as the Deputy Editor-in-Chief for *IET Intelligent Transport Systems* journal and an Associate Editor for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE/ASME TRANSACTIONS ON MECHATRONICS, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE/CAA JOURNAL OF AUTOMATICA SINICA, IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, *Journal of Dynamic Systems* (ASME), and *Measurement and Control*. He was a Guest Editor of *Vehicle System Dynamics*, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, and IEEE INTERNET OF THINGS JOURNAL.