V2VFormer: Vehicle-to-Vehicle Cooperative Perception with Spatial-Channel Transformer

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

***Abstract*—Collaborative perception aims for a holistic perceptive construction by leveraging complementary information from nearby connected automated vehicle (CAV), thereby endowing the broader probing scope. Nonetheless, how to aggregate individual observation reasonably remains an open problem. In this paper, we propose a novel vehicle-to-vehicle perception framework dubbed *V2VFormer* with *T*ransformer-based *Co*llaboration (*CoTr*). Specifically. it re-calibrates feature importance according to position correlation via Spatial-Aware Transformer (*SAT*), and then performs dynamic semantic interaction with Channel-Wise Transformer (*CWT*). Of note, *CoTr* is a light-weight and plug-in-play module that can be adapted seamlessly to the off-the-shelf 3D detectors with an acceptable computational overhead. Additionally, a large-scale cooperative perception dataset V2V-Set is further augmented with a variety of driving conditions, thereby providing extensive knowledge for model pretraining. Qualitative and quantitative experiments demonstrate our proposed *V2VFormer* achieves the state-of-the-art (SOTA) collaboration performance in both simulated and real-world scenarios, outperforming all counterparts by a substantial margin. We expect this would propel the progress of networked autonomous-driving research in the future.**

***Index Terms*—Vehicle-to-Vehicle (V2V) Collaboration, Cooperative Perception, Autonomous Driving, Transformer, Intelligent Transportation Systems.**

I. Introduction

INtelligent transportation systems (ITS) has demonstrated a tremendous potential to facilitate the safety and efficiency of traffic operation [1]. As an essential ingredient, environmental perception provides an adequate comprehension of surroundings and participants from heterogeneous sensory data, thus gaining widespread popularity. The emergence in deep learning pushes a remarkable step for self-driving perception, and its accuracy/robustness has been significantly improved in several tasks such as object detection [2] [3] [4] [5], multi-object tracking [6], segmentation [7], etc. Despite its immense potential, single-agent perception suffers from occlusion, blind spot and sparse measurement (*i.e.*, LiDAR point) challenges, and individual perspective easily causes an unreliable and uncertainty prediction especially in presence of severe occlusions and corner cases (*e.g.*, sensor failure), as depicted in Fig.1. Such problem would cause a catastrophic accident, thereby posing a threat to driving safety. To address this, recent endeavors are greatly dedicated to explore a collaborative perception paradigm [8] [9] [10] [11] where sensory information from nearby CAVs could be transmitted and shared with each other via vehicle-to-vehicle (V2V) communication [12], consequently preventing the ego vehicle from unknown hazard in the real world.

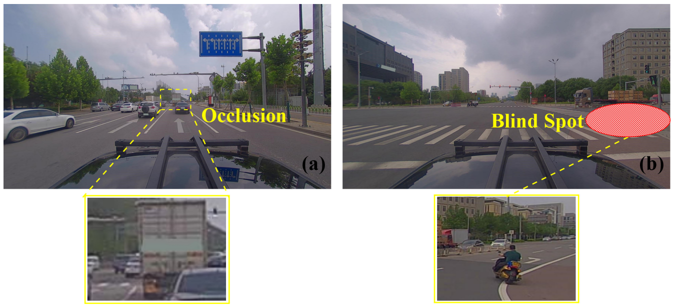


Fig. 1. The prototype of individual perception in real-world scenario [13]. ***Left.*** Vehicle easily suffers from occlusion in ego-centric view, but it can be mitigated by complementary observations from networked cars in the vicinity. ***Right.*** Due to the blind-spot area, ego-vehicle fails to foresee the rear-side participant when turning right at an intersection, possibly causing a severe traffic conflict or accident-prone hazard.

The core issue of cooperative perception is to determine what information should be received from nearby agents and how to collaborate valuable information for perception enhancement. Based on various fusion approaches, contemporary works can be roughly divided into three categories. Early collaboration [8] incorporates raw sensory measurement among agents, and circumvent occluded or blind-spot risk via a global viewpoint. Whereas, redundant information would introduce unaffordable computation budget and transmission bandwidth. Late collaboration is a communication-efficient pipeline of combining the independent prediction directly, while accumulated error from single-vehicle anticipation would result in an unsatisfactory performance unavoidably. Intermediate collaboration [9] [14] has been viewed as a communication-performance trade-off paradigm where informative feature is aggregated from neighboring cars to enhance perception ability. Nevertheless, these algorithms mainly depend on hard-attention association with scalar-valued calculation, making spatial correlation and information interaction not being well-studied. It is consequently non-trivial to develop an adaptative and powerful collaboration for a sufficient exploration of

Corresponding anthor: Daxin Tian (dtian@buaa.edu.cn).

C. Lin, D. Tian, X. Duan, J. Zhou are with the State Key Laboratory of Intelligent Transportation System, Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control, Zhongguancun Laboratory, School of Transportation Science and Engineering, Beihang University, Beijing 100191, China.

D. Zhao is with the James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

D. Cao is with the Department of Mechanical and Mechatronics Engineering, University of Waterloo, 200 University Ave West, Waterloo ON, N2L3G1 Canada

multi-agent advantageous observations.

Furthermore, constructing an extensive cooperative perception dataset is also imperative to model training and evaluation. There is a spectrum of benchmarks publicly-available for single-vehicle perception, *e.g.*, KITTI [15], nuScenes [16], Waymo [17], *etc.*, with synchronous calibration and fine-grained annotation. In the context of multi-agent perception, it is of particular interest for data simulation due to costly acquisition and labeling [18]. For instance, V2V-Sim [14] builds a group of new scenarios from the real-world collections based on a self-driving LiDAR simulation [19], and however, both of them are not released. On the top of OpenCDA [20] platform co-simulated with CARLA [21] and SUMO [22], OPV2V [23] customizes a V2V communication dataset with over $10k$ frame RGB images and LiDAR points. Chen et al. [24] develops a data collection system that integrates LiDAR point cloud from both CAVs and infrastructure for various transportation applications. Moreover, V2V4Real [25] is the first large-scale realistic V2V perception benchmark for practical deployment.

In this paper, we recast collaborative perception into a multi-vehicle LiDAR-based 3D detection task, and introduce a novel vehicle-to-vehicle cooperative perception framework termed as *V2VFormer*, that is composed of data sharing and extraction, feature compression and fusion, and prediction header. After relative pose and extrinsic information sharing within a communication range (*i.e.*, $70m$ [12]), intermediate maps from nearby CAVs are projected and transformed into the ego-vehicle coordinate, respectively. By leveraging the advantage of transformer in both spatial and channel feature learning, we further propose a *Tr*ansformer-based *Co*llaboration dubbed *CoTr* with Spatial-Aware Transformer (*SAT*) and Channel-Wise Transformer (*CWT*) for intermediate aggregation. The former highlights the potential foreground/target region according to multi-agent location correlation, while the latter is responsible for channel-wise interaction across each paired ego-networked vehicles. We claim that *CoTr* is a light-weight and play-in-plug module with considerable linear complexity and flexible adaptable for various individual perception architectures. Finally, two branches with feed forward network (FFN) are utilized for predicting classification confidence and box regression, respectively. Moreover, a newly-built cooperative perception dataset V2V-Set is further introduced under sensor suites and configurable settings of OPV2V [23], which provides a variety of scenarios with challenging conditions (e.g., changing weathers and illuminations) for model pretraining. Qualitative and quantitative experiments are extensively conducted on OPV2V, V2X-Sim and V2V4Real datasets, and our proposed *V2VFormer* achieves the state-of-the-art 3D detection accuracy over both individual and collaborative counterparts by a remarkable margin, demonstrating its superiority and generalization in both simulation and reality.

In general, our main contributions can be concluded as follows:

- We introduce a novel vehicle-to-vehicle cooperative 3D object detection paradigm named as *V2VFormer*, that consists of data sharing and extraction, feature transmission and fusion, and prediction header.
- To enhance spatial-channel information exchange across networked vehicles in the vicinity, *Tr*ansformer-based *Co*llaboration (*CoTr*) is designed with Spatial-Aware Transformer (*SAT*) for spatial relation encoding among nearby agents and Channel-Wise Transformer (*CWT*) for semantic feature interaction in an adaptative manner.
- On the top of open-sourced OPV2V benchmark, a large-scale cooperative perception dataset V2V-Set is further constructed with diverse driving conditions for supporting model pretraining with abundant priors.
- Empirical studies on OPV2V, V2X-Sim and V2V4Real benchmarks consistently demonstrates the effectiveness and advancement of *V2VFormer*, which outperforms both single-agent and multi-agent alternatives by a distinct margin.

The reminder in this paper is organized as follows: we review related works in Section II, and describe our proposed method in Section III. Section IV presents implementation setup and experimental result, and we conclude the overall research in Section V.

II. Related Works

This section overviews the contemporary development on LiDAR-based 3D detection, vehicle-to-vehicle perception and datasets briefly.

A. LiDAR-based 3D Detection

LiDAR-based 3D object detection has been widely explored based on different data formats, *e.g.*, raw point [27] [28], voxel grid [29] [30] [31] and hybrid representation [32] [33]. Based on PointNet [34], PointRCNN [35] pioneers a point-based 3D detection architecture that performs foreground segmentation for box estimation at first and conducts proposal refinement with semantic attributes later. To improve sampling strategy, 3DSSD [36] incorporates distance- and feature-farthest point sampling to generate candidate points, while Chen et al. [37] develop semantics-augmented set abstraction (SASA) to retain more important foreground point. VoxelNet [26] quantizes point cloud into equally-spaced 3D volumetric grid and transforms a group of points within each voxel into a unified representation through voxel feature encoding (VFE) layer. To speed up quantization efficiency, Yan et al. [38] designs 3D sparse convolutional neural network (CNN) for accelerating voxel feature learning, and PointPillars [39] converts raw point into a pseudo-image and utilizes standard 2D convolution for downstream tasks. Moreover, point-voxel hybrid representation is beneficial to alleviate grid quantization loss and boost 3D detection accuracy. PV-RCNN [40] represents the 3D scene as a small set of representative points, and transforms proposal-specific features into a RoI-grid points via keypoint set abstraction. Recent transformer-based 3D detection [41] [42] is drawn increasing attention from research community. Pointformer [43] is a pure transformer backbone for point-based 3D detection, that contains local, local-global and global transformers to integrates both context-dependent and context-aware features at multiple resolutions. Mao et al. [44] utilize local attention and dilated attention for operating empty and

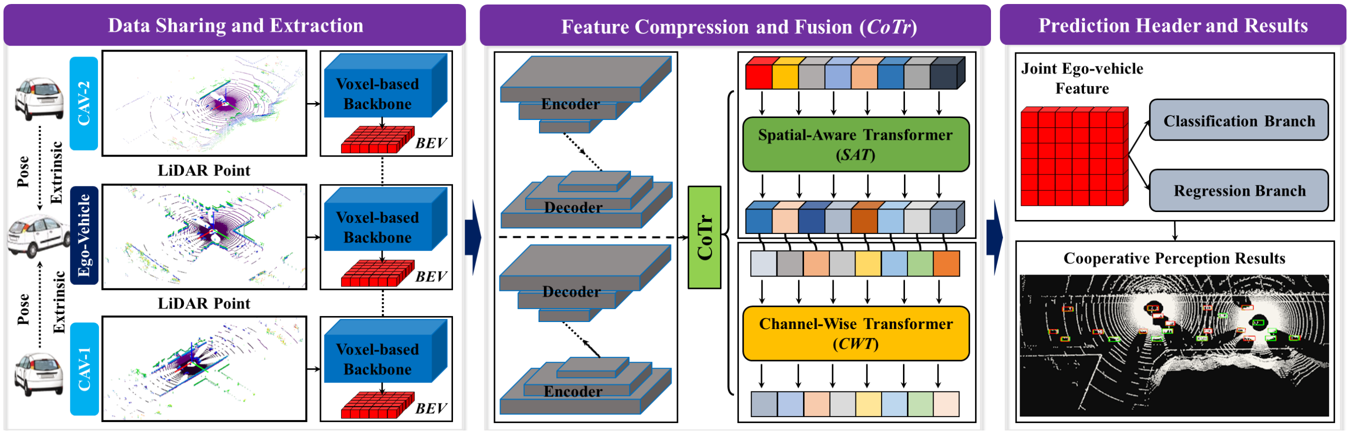
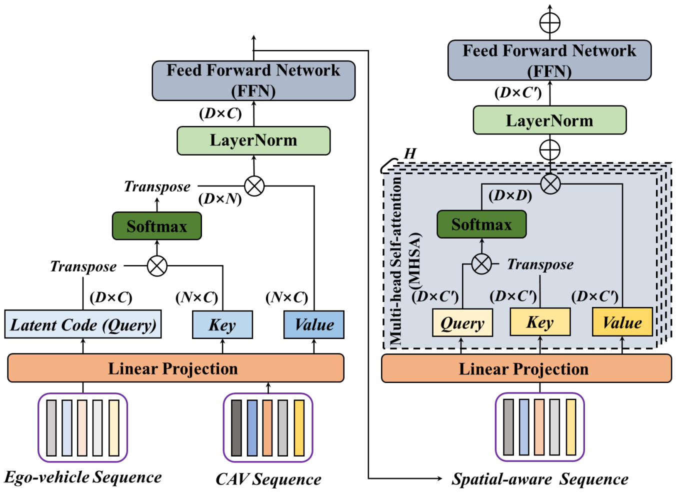

Data Sharing and Extraction
Feature Compression and Fusion (CoTr)
Prediction Header and Results
CAV-2
Ego-Vehicle
CAV-1
Pose
Extrinsic
Pose
Extrinsic
LiDAR Point
LiDAR Point
Voxel-based Backbone
Voxel-based Backbone
Voxel-based Backbone
BEV
BEV
BEV
Encoder
Decoder
Decoder
Encoder
CoTr
Spatial-Aware Transformer (SAT)
Channel-Wise Transformer (CWT)
Joint Ego-vehicle Feature
Classification Branch
Regression Branch
Cooperative Perception Results

Fig. 2. The overview of ***V2VFormer*** pipeline. **(a) Data Sharing and Extraction:** LiDAR point from independent vehicle is projected on the ego coordinate according to sharing pose and extrinsic parameters, and voxel-based feature backbone (e.g., VoxelNet [26]) is then utilized for bird-eye's view (BEV) map generation at each single perspective. **(b) Feature Compression and Fusion (*CoTr*):** an encoder-decoder network with 1×1 convolutions is viewed as feature compression for transmission bandwidth saving. Given a paired ego-CAV map, *T*ransformer-based *Co*llaboration dubbed *CoTr* is designed with Spatial-Aware Transformer (*SAT*) for intermediate feature re-calibration according to the geometry correlation, and Channel-Wise Transformer (*CWT*) for detailed semantics interaction across agents. **(c) Prediction Header and Results:** with the joint feature map, two branches with feed forward network (FFN) is proposed for classification confidence prediction and bounding-box regression. The cooperative perception result is visualized with prediction (*Red*) and ground-truth (GT) (*Green*). Best viewed in color.

non-empty voxel position, respectively. In this paper, we formulate multi-agent perception problem as LiDAR-only 3D object detection problem.

B. Vehicle-to-Vehicle Perception

The goal of recently-emerged cooperative perception [45] [46] [47] [48] [49] is to upgrade ego-vehicle recognition ability by receiving the messages from neighboring CAVs within an acceptable vehicle-to-vehicle communication range (i.e., $70m$ [12]). Prior works can be roughly classified into early, intermediate and late collaborations. Cooper [8] firstly conducts raw-data cooperative perception by combining LiDAR point cloud from different positions and angles of connected vehicles. However, sharing raw point without partitioning is restricted by available communication bandwidth and computational overhead. Contrarily, late fusion is explored by combining the independent detection from spatially-diverse sensors [10]. Albeit less communication bandwidth, it severely damages the perception accuracy due to the devoid of context and false-positive prediction. Intuitively, intermediate collaboration could be performance-efficiency trade-off pipeline that aggregates collective features across agents in the vicinity. V2VNet [14] updates node message via mask-aware permutation-invariant function and convolutional gated recurrent unit (ConvGRU). DiscoNet [50] is a brand-new knowledge distillation framework where teacher model employs an early combination with holistic-view inputs, and student network conducts intermediate fusion with single-view feature. Recently, the first open-sourced V2V simulated benchmark OPV2V [23] stimulates a spectrum of cooperative perception algorithms: V2X-ViT [51] designs heterogeneous multi-agent and multi-scale window self-attention modules for information fusion across on-road vehicles and infrastructures. Meng et al. [47] incorporates historical tracking cue explicitly via spatial-temporal 3D network. CORE [52] instantiates multi-agent perception with a compressor, an attentive collaboration and a reconstruction module in a learning-to-reconstruct manner. Nevertheless, it is unreasonable to deal with the contribution of CAVs as equal, thereby causing key-point feature at different positions not be fully leveraged. To this end, we propose a novel V2V cooperative perception with *T*ransformer-based *Co*llaboration (*CoTr*) for both spatial-channel awareness.



(a). Spatial-Aware Transformer (*SAT*)

(b). Channel-Wise Transformer (*CWT*)

Fig. 3. The schematic illustration of ***Spatial-Aware Transformer (SAT)*** and ***Channel-Wise Transformer (CWT)*** modules. Noted that ⊕ and ⊗ denote element-wise concatenation and matrix multiplication, respectively.

C. Vehicle-to-Vehicle Datasets

Previous works involve several vehicle-to-vehicle datasets by virtue of available single-agent data [53] or autonomous-driving simulation [54]. On the one hand, Chen et al [1] [24] select various frames from KITTI [15] to configure multi-vehicle setting at different timestamps, but spatial offset and viewpoint overlapping among cars would be unavoidable. On the other hand, V2V-Sim [14] synthesizes the unseen driving scenario on the basis of real-world collections, while it is not publicly-available. Research community demands a large-scale and widespread cooperative perception benchmark for comprehensive model training and evaluation, and OPV2V [23] is therefore constructed with more than $10k$ frames over 70 intersection situations from 8 digital towns. More importantly, it provides a suite of sensor configuration and setting [21], making it reproducible for both academia and industry. V2X-Sim [50] is composed of 100 scenes in total of $10k$ frames for LiDAR-based vehicle-to-vehicle algorithm evaluation. Recent interest on real-world cooperative perception facilitates the development of V2V4Real [25], that is the first large-scale real-world multi-modal dataset for V2V perception comprising $200k$ LiDAR frames, $40k$ RGB frames and $240k$ annotated 3D bounding boxes for 5 classes. In this work, a new-built V2V-Set is launched on the top of OPV2V with a variety of challenging conditions, and empirical study is conducted to investigate its cooperative performance on both simulation and reality.

III. Methodology

This section revisits the preliminary of cooperative perception and vision transformer, and technical details of vehicle-to-vehicle perception pipeline *V2VFormer* with *Tr*ansformer-based *Co*llaboration (*CoTr*) is introduced later.

A. Prerequisites

Cooperative Perception. It is formulated as multi-agent LiDAR-based 3D object detection issue where a target vehicle strategically receives observations from nearby CAVs within a vehicle-to-vehicle communication range (*i.e.*, $70m$ [12]), aiming for longer probing range and broader perception scope. Formally, an ego vehicle V_e is selected from a group of connected vehicles V_m $(m = 1, \cdots, M)$ at timestamp t, where M implies the number of CAVs in the vicinity. For simplicity, we assume each agent provides an accurate spatial location with well-synchronized LiDAR measurement, denoted by $P_{V_m} \in \mathbb{R}^{N\times 3}$. Subsequently, individual map F_{V_m} with a shape of $H \times W \times 3$ is incorporated at the ego-centric coordinate, and we obtain box prediction $\mathcal{B}$ with category score $\mathcal{S}$ for each object, where $\mathcal{B}$ contains box center coordinate (x, y, z), spatial size (l, w, h) and orientation θ.

Vision Transformer. Self-attention (SA) is the core component in the vision transformer [55] [56] [57], which attends to the discriminative object region via matrix multiplication between Query (Q), Key (K) and Value (V) embeddings. Given an input feature $X \in \mathbb{R}^{L\times C}$ of sequence length L and channel number C, SA can be mathematically described as Eq.1 and Eq.2:

$$Q = W_Q X, K = W_K X, V = W_V X \tag{1}$$

$$SA(Q, K, V) = \sigma(\frac{QK^T}{\sqrt{d}}) \bigotimes V \tag{2}$$

where $W_* \in \mathbb{R}^{L\times L}$ are learnable linearity weights, respectively, d denotes head dimension, $\sigma(\cdot)$ is softmax normalization and $\bigotimes$ is element-wise matrix multiplication. Due to its quadratic computational overhead, employing self-attention calculation directly on large-scale raw point or low-level feature map is prohibitive for limited communication resource.

The findings of SA approximated with low-rank projection makes it apt for a k-clustering process where hidden vector is served as cluster center [42] [58]. Therefore, a standard SA could be decomposed into two cross-attention (CA) operators induced by a group of *latent code* $L \in \mathbb{R}^{K\times C}$ with size K, as formulated in Eq.3 and Eq.4:

$$SA(L, \mathcal{X}) = CA(\mathcal{X}, FFN(CA(L, \mathcal{X}))) \tag{3}$$

$$CA(L, \mathcal{X}) = \sigma(L^T \mathcal{X}) \bigotimes \mathcal{X} \tag{4}$$

where $\mathcal{X} \in \mathbb{R}^{L\times C}$ refers to a projection of input X (*i.e.*, linear transformation), and $FFN(\cdot)$ implies feed forward network. Generally, CA could distill the most significant ingredient of embedding at first, and a set of *latent code* is further highlighted for cross-attention interaction. More importantly, CA is feasible and affordable due to its linear complexity only to input length N and *latent code* number K.

B. V2VFormer: Vehicle-to-Vehicle Cooperative Perception

The overall architecture of *V2VFormer* includes data sharing and extraction, feature compression and fusion, and prediction header modules, as depicted in Fig.2.

1) Data Sharing and Extraction: as commonly done in previous works [23], a spatial graph is constructed based on relative pose and extrinsic of CAVs, each node of which represents a networked vehicle within the communication range. Individual LiDAR is then projected into a unified reference plane (*i.e.*, ego coordinate) for data alignment, and finally, we adopt voxel-based feature backbone for efficiency, *e.g.*, VoxelNet [26], SECOND [38] and PointPillars [39], producing an intermediate bird's-eye view (BEV) map $F_{V_m} \in \mathbb{R}^{H\times W\times C}$.

2) Feature Compression and Fusion: an encoder-decoder compression module is introduced for mitigating expensive transmission bandwidth: the encoder stacks several 1×1 convolutions for BEV feature compression progressively along channel dimension, and the decoder projects it back into the original feature resolution via corresponding 1×1 deconvolutions. Afterwards, *Tr*ansformer-based *Co*llaboration (*CoTr*) is designed for intermediate feature combination from nearby CAVs.

CoTr. Given an ego-centric map $F_{V_m} \in \mathbb{R}^{H\times W\times C}$, an intuitive solution is to directly incorporate different representations among vehicles by a simple operator (*i.e.*, addition or summation). Nonetheless, geometry correlation and semantic interaction across agents could not be fully exploited in a hard-association manner. To this end, a novel *Tr*ansformer-based *Co*llaboration (*CoTr*) is proposed for both spatial-channel awareness. As demonstrated in Fig.3, Spatial-Aware Transformer (*SAT*) is responsible for agent importance re-calibration according to spatial distance via cross-attention operation,

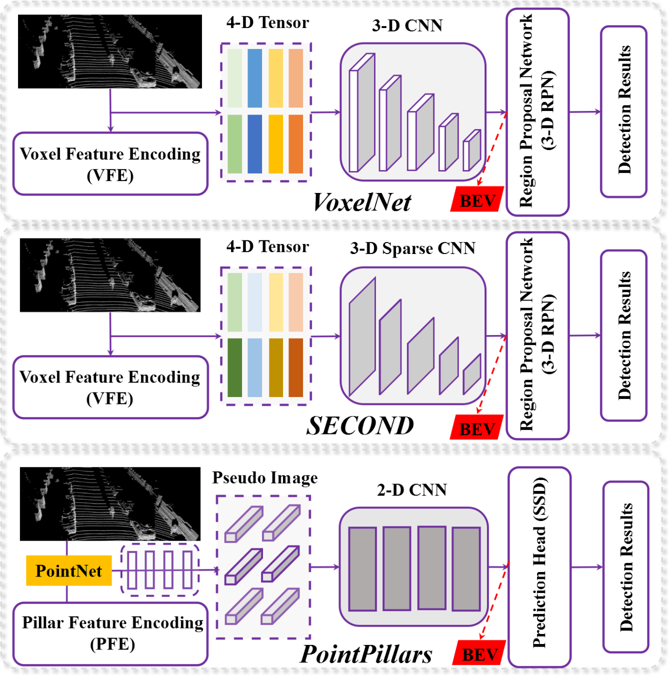


Fig. 4. The pipeline of three voxel-based 3D object detector, *i.e.*, ***VoxelNet*** [26], ***SECOND*** [38] and ***Pointpillars*** [39]. Compared to VoxelNet, SECOND improves a 3D sparse convolutional neural network (3-D Sparse CNN) for efficiency, while pillar feature encoding (PFE) in PointPillars compresses voxel grid into a 2-D pseudo image along the vertical column.

and Channel-Wise Transformer (*CWT*) further conducts Multi-Head Self-Attention (MHSA) for channel semantic interaction among each ego-networked pair. In this way, ego-centric map could be sufficiently reinforced with both spatial and semantic information, and we introduce technical details step-by-step.

a. Spatial-Aware Transformer (*SAT*) It determines which agent should be collaborated according to spatial distance. As illustrated in Fig.3-(a), $F_{V_m} \in \mathbb{R}^{H\times W\times C}$ is firstly flattened and mapped into the hidden feature space using a linear projection, resulting in a sequential vector $\hat{F}_{V_m} \in \mathbb{R}^{N\times C}(N = HW)$. For each pair of ego and i-th connected vehicles $[V_e, V_m]$ $(m \neq e)$, a fixed-length *latent code* $L_m \in \mathbb{R}^{D\times C}(D = \frac{N}{M})$ is abstracted from ego input $\hat{F}_{V_e}$, and the other CAV feature is transformed into Key $K_m^S \in \mathbb{R}^{N\times C}$ and Value $V_m^S \in \mathbb{R}^{N\times C}$ embeddings with linear projection, respectively.

The cross-attention calculation is then launched between latent vector and key tensor, and after a normalized softmax function $\sigma(\cdot)$, attention matrix $\mathcal{A}_m \in \mathbb{R}^{N\times D}$ is obtained to indicate the importance score of ego-networked vehicle being aware of spatial correlation. Finally, we generate spatial-aware feature vector $F_{(V_e\leftarrow V_m)} \in \mathbb{R}^{D\times C}$ by incorporating K/V embeddings with attention matrix $\mathcal{A}_m$ in an element-wise multiplication, as mathematically concluded in Eq.5∼Eq.7.

$$\begin{aligned} K_m^S : \mathcal{F}^1_{V_m} &= W^S_K \hat{F}_{V_m} \\ V_m^S : \mathcal{F}^2_{V_m} &= W^S_V \hat{F}_{V_m} \end{aligned} \tag{5}$$

$$\mathcal{A}_m(L_m, \mathcal{F}^1_{V_m}) = \sigma(\mathcal{F}^1_{V_m} L_m^T) \tag{6}$$

$$\begin{aligned} F_{(V_e\leftarrow V_m)} =& CA(L_m, K_m^S, V_m^S) \\ =& \mathcal{A}_m^T(L_m, \mathcal{F}^1_{V_m}) \bigotimes \mathcal{F}^2_{V_m} \end{aligned} \tag{7}$$

b. Channel-Wise Transformer (CWT) A naive scheme is to concatenate the encoded sequence along length dimension to recover the spatial resolution, while the introduction of feature redundancy or background noise would be detrimental to perception accuracy. Hence, Channel-Wise Transformer (*CWT*) is further designed for each ego-CAV semantic interaction across channels. As presented in Fig.3-(b), a group of Query (Q_m^C), Key (K_m^C) and Value (V_m^C) is firstly transformed from $F_{(V_e\leftarrow V_m)} \in \mathbb{R}^{D\times C}$, via a linear projection embedded with positional information. It is noted that position encoding is great of essential to indicate spatial location for each ego-networked vehicle. Subsequently, channel attention is obtained by normalizing Multi-Head Self-Attention (MHSA) output with a softmax function, and we produce channel-wise feature sequence $\tilde{F}_{(V_e\leftarrow V_m)} \in \mathbb{R}^{D\times \tilde{C}}$ by matrix multiplication, where $\tilde{C} = C$ denotes the interacted channel number. Finally, we concatenate them along the spatial dimension, forming a joint ego-vehicle sequence $\tilde{F}_{V_e} \in \mathbb{R}^{N\times \tilde{C}}$. The whole process can be formulated as Eq.8∼Eq.10: *Concate*$[\cdot]$ defines element-wise concatenation; H is head number (*i.e.*, 16).

$$\begin{aligned} Q_m^C &= W_Q^C F_{(V_e\leftarrow V_m)}, \\ K_m^C &= W_K^C F_{(V_e\leftarrow V_m)}, \\ V_m^C &= W_V^C F_{(V_e\leftarrow V_m)} \end{aligned} \tag{8}$$

$$\begin{aligned} \tilde{F}_{(V_e\leftarrow V_m)} =& Concate[SA_h(Q_m^C, K_m^C, V_m^C)] \\ =& Concate[\sigma(\frac{Q_m^C (K_m^C)^T}{\sqrt{d_m}}) \bigotimes V_m^C] \\ & h = 1, \cdots, H \end{aligned} \tag{9}$$

$$\tilde{F}_{V_e} = Concate[\tilde{F}_{(V_e\leftarrow V_m)}], m = 1, \cdots, M \tag{10}$$

In general, there are several advantages of *CoTr* strategy for multi-vehicle feature aggregation. On the one hand, Spatial-Aware Transformer (*SAT*) re-calibrates the feature significance of nearby CAVs according to spatial correlation, thereby attending to foreground target more. On the other hand, Channel-Wise Transformer (*CWT*) further underlines detailed semantic interaction across agent channels in a dynamic manner. Compared to other alternatives [51] [47] [52] [59], our *CoTr* performs both spatial-channel awareness among per-agent, which could promote for better multi-agent collaboration and perception ability in the challenging scenarios. More importantly, linear complexity of *CoTr* $[\mathcal{O}(NDC) + \mathcal{O}(C^2)]$ $(C \ll N)$ demonstrates its efficiency and feasibility in other perception framework with acceptable overhead.

3) Prediction Header: The fused ego feature $\tilde{F}_{V_e} \in \mathbb{R}^{N\times \tilde{C}}$ is recovered back into a format of 2D BEV map $\tilde{F} \in \mathbb{R}^{H\times W\times \tilde{C}}$ via tensor reshape operator, and we employ two feed forward network (FFN) branches for object classification and 3D box regression, respectively.

TABLE I

Performance comparisons with different collaborations achieved by VoxelNet [26], SECOND [38] and PointPillars [39] on **OPV2V *Default*** split. The model pre-trained on OPV2V and V2V-Set reports 3D AP on the both sides of sign '/', and we highlight the best result with bold font.

Methods	**OPV2V *Default***							
	AP@50 (%)				AP@70 (%)			
	No	*Early*	*Late*	*CoTr*	*No*	*Early*	*Late*	*CoTr*
VoxelNet	68.8/69.2	75.8/76.7	73.8/74.0	86.3/**87.4**	52.6/53.1	67.7/67.8	58.8/59.2	81.1/**81.3**
SECOND	71.3/71.9	81.3/83.2	77.5/78.0	91.7/**92.1**	60.4/61.6	74.1/75.0	70.1/70.8	87.6/**88.0**
PointPillars	67.9/68.1	80.0/80.9	78.1/78.7	88.6/**89.2**	60.2/61.0	69.6/71.1	66.8/67.0	82.8/**84.4**

TABLE II

Performance comparisons with different collaborations achieved by VoxelNet [26], SECOND [38] and PointPillars [39] on **OPV2V *Culver City*** split. The model pre-trained on OPV2V and V2V-Set reports 3D AP on the both sides of sign '/', and we highlight the best result with bold font.

Methods	**OPV2V *Culver City***							
	AP@50 (%)				AP@70 (%)			
	No	*Early*	*Late*	*CoTr*	*No*	*Early*	*Late*	*CoTr*
VoxelNet	60.5/60.9	67.7/68.0	58.8/58.7	80.7/**81.5**	43.1/43.6	61.3/62.1	50.8/52.3	75.2/**76.1**
SECOND	64.6/65.5	73.8/74.6	68.2/69.1	89.3/**89.8**	51.7/52.0	66.4/66.7	60.2/60.9	80.5/**81.2**
PointPillars	55.7/56.3	69.6/70.1	66.8/67.5	83.8/**84.6**	47.1/47.3	62.2/62.7	60.1/60.4	78.5/**79.8**

C. Loss Function

The overall loss function $\mathcal{L}$ is formulated as classification loss $\mathcal{L}_{cls}$, regression loss $\mathcal{L}_{reg}$ as well as direction loss $\mathcal{L}_{dir}$:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{dir} \tag{11}$$

To be specific, classification loss $\mathcal{L}_{cls}$ utilizes focal loss [60] to alleviate foreground-background imbalance, while both regression loss $\mathcal{L}_{reg}$ and direction loss $\mathcal{L}_{dir}$ adopt the smooth-$L1$ loss. The former calculates the localization offset ($\triangle x, \triangle y, \triangle z, \triangle l, \triangle w, \triangle h$), and the latter utilizes *Sinusoidal* function for orientation residual $\triangle\theta = \sin(\theta^{gt} - \theta^{p})$ between ground-truth (GT) and prediction. More details can further refer to [26] [38] [39].

IV. Experiments

This section firstly introduces a newly-built cooperative perception dataset V2V-Set on the basis of OPV2V. And then we conduct extensive experiments and ablation study on both simulated and real-world benchmarks, to verify the effectiveness and contribution of proposed method.

A. Datasets

OPV2V/V2V-Set. OPV2V [23] pioneers a publicly-available vehicle-to-vehicle perception benchmark co-simulated with CARLA [21] and SUMO [22], which includes 11464 image and point data with 232913 annotations covering more than 70 scenes from 8 towns on CARLA map.

To further augment the variety of OPV2V, V2V-Set is constructed with changing weather (sunny, rainy and cloudy) and illumination (dawn, dusk and nighttime) conditions. It is acquired with four cameras providing 360° viewpoint and 64-channel LiDAR mounted on the vehicle. In total, V2V-Set consists of over 100 scenarios with 20000 frames, each of which contains 2 to 8 CAVs. It is argued that a diversity of training examples could offer rich knowledge for performance increment, and experimental result would verify this view.

V2X-Sim. V2X-Sim covers LiDAR-based vehicle-to-vehicle scenarios at a certain intersection of *Town05* generated by CARLA and SUMO platforms. In general, it contains 100 scenes with a total of 10000 samples. Each scene consists of 100 frames with a 20-second traffic flow, and $2 \sim 5$ vehicles are selected as collaborative agents.

V2V4Real. V2V4Real is the first large-scale real-world dataset for V2V perception, which comprises $20k$ LiDAR and $40k$ image frames with $240k$ 3D labeling for 5 categories. It is challenged that objects have a diversity of box sizes with length ranging from $2.5m$ to $23m$, widths ranging from $1.5m$ to $4.5m$ and heights ranging from $1m$ to $4.5m$, respectively.

B. Implementation Details

As illustrated in Fig.4, we adopt **VoxelNet** [26], **SECOND** [38] and **PointPillars** [39] as backbone for high-efficiency voxel-based representation learning, and incorporate them with different fusion strategies for a thorough analysis: *no collaboration* implies single-agent perception without any received information; *early collaboration* aggregates the raw LiDAR projected on egocentric coordinate for feature extraction; *late collaboration* combines independent prediction from each CAV, and produces the final result via post-processing Non-Maximum Suppression (NMS) operator.

Unless otherwise specified, the range of point cloud is set to $[-140, 140]m$, $[-40, 40]m$ and $[-3, 1]m$ for both OPV2V and

TABLE III
PERFORMANCE COMPARISONS WITH SOTA METHODS ON **OPV2V *Default*/*Culver City*** SPLITS AND **V2X-SIM *Test*** SET. THE BEST RESULT IS HIGHLIGHTED WITH BOLD FONT, AND WE DRAW THE ACCURACY INCREMENT WITH RED COLOR IN THE BRACKET.

Methods	**OPV2V *Default***		**OPV2V *Culver City***		**V2X-Sim *Test***	
	AP@50 (%)	AP@70 (%)	AP@50 (%)	AP@70 (%)	AP@50 (%)	AP@70 (%)
V2VNet [14]	82.2	79.4	73.4	68.9	56.8	50.7
DiscoNet [50]	80.1	76.8	75.0	70.3	*60.3*	*53.9*
OPV2V-attn [23]	86.4	80.2	77.5	73.6	59.4	53.3
CoBEVT [59]	86.1	81.6	77.3	73.1	-	-
FPV-RCNN [61]	82.0	77.3	76.3	72.0	-	-
Where2comm [62]	88.9	75.5	82.2	68.0	59.1	52.2
CORE [52]	*90.9*	*85.8*	*87.7*	*78.1*	-	-
SECOND-*CoTr*	**91.7** (**+0.8**)	**87.6** (**+1.8**)	**89.3** (**+1.6**)	**80.5** (**+2.4**)	**61.5** (**+1.2**)	**55.0** (**+1.1**)

V2V4Real, and $[-32, 32]m$, $[-32, 32]m$, $[-3, 2]m$ for V2X-Sim along x-y-z axes. More details about network parameters and training settings could refer to the open-sourced *OpenCOOD*[1], *Coperception*[2] and *V2V4Real*[3], respectively. All experiments are conducted on Ubuntu20.04 with two NVIDIA RTX3090 GPUs. Moreover, we split 6764/1981/2719 in OPV2V, 16000/4000 in V2V-Set, 8000/900/1100 in V2X-Sim, 14210/2000/3986 in V2V4Real for model training/validation/testing, and report average precision (AP) at 0.5/0.7 intersection-of-union (IoU) thresholds for comparison.

TABLE IV
PERFORMANCE COMPARISONS WITH SOTA METHODS ON **V2V4REAL *Test***. WE HIGHLIGHT THE BEST RESULT WITH BOLD FONT.

Methods	**V2V4Real *Test***	
	AP@50 (%)	AP@70 (%)
No Collaboration	39.8	22.0
Early Collaboration	59.7	32.1
Late Collaboration	55.0	26.7
V2VNet [14]	64.7	33.6
OPV2V-attn [23]	64.5	34.3
V2V-ViT [51]	64.9	**36.9**
CoBEVT [59]	**66.5**	36.0
PointPillars-*CoTr*	*65.4*	*36.1*

C. Evaluation Results

Comparisons with Different Collaborations. Table I and Table II elaborate the detection results of VoxelNet [26], SECOND [38] and PointPillars [39] with different fusions on OPV2V *Default* and *Culver City* test splits in terms of 0.5/0.7 IoU thresholds, respectively. Taking an example of collaboration perception on *Default* set, it is clearly observed that intermediate collaboration (*CoTr*) reports the best AP@50 of 86.3%/91.7%/88.7% and AP@70 of 81.1%/87.6%/82.8% when incorporated with three different backbones, which surpasses the counterparts over a remarkable margin. For instance, SECOND-*CoTr* improves *No*/*Early*/*Late* collaborations by 20.4%/10.4%/14.2% AP@50 and 27.2%/13.5%/17.5% AP@70, while Pointpillars-*CoTr* establishes 8.6% $\sim$ 20.7% and 13.2% $\sim$ 22.6% AP gains at both IoU thresholds. More importantly, an non-negligible accuracy upgrade, *e.g.*, 0.4% $\sim$ 1.6% averaging growth, could be further received when model pre-trained on V2V-set, which implies the advantage of large-scale training data. Similar improvements on *Culver City* test are also obtained from Table II.

Due to both spatial-aware and channel-wise feature being fully exploited, it is reasonably argued that *CoTr* strategy contributes to a better collaboration performance. Moreover, a variety of driving scenarios in V2V-Set provides an essential prior for meaningful and versatile knowledge learning, thereby facilitating a higher perception accuracy. The consistent improvement confirms its effectiveness and contribution.

TABLE V
ABLATION STUDY ON THE EFFECTIVENESS OF ***Collaborative Strategy*** ON ***OPV2V Default*** SPLIT. THE BEST RESULT IS HIGHLIGHTED WITH BOLD FONT, AND WE DRAW THE ACCURACY INCREMENT WITH RED COLOR IN THE BRACKET.

Collaborations	**OPV2V *Default***	
	AP@50 (%)	AP@70 (%)
PointPillars	67.9	60.2
w./Average	76.0 (+8.1)	67.2 (+7.0)
w./Max	74.9 (+7.0)	65.3 (+3.9)
w./Mask-aware	82.2 (+14.3)	79.4 (+19.2)
w./Attentive	86.4 (+18.5)	80.2 (+20.0)
PointPillars-*CoTr*	**88.6** (+20.7)	**82.8** (+22.6)

Comparisons with SOTA Methods. To comprehensively validate the superiority of *CoTr*, empirical analysis is further conducted on OPV2V, V2X-Sim and V2V4Real benchmarks in comparisons of the state-of-the-art (SOTA) methods, and we choose SECOND backbone with CoTr scheme for simplicity.

As shown in Table III, SECOND-*CoTr* sets a new state-of-the-art on three test splits and outperforms all counterparts

[1] https://github.com/DerrickXuNu/OpenCOOD
[2] https://github.com/coperception/coperception
[3] https://github.com/ucla-mobility/V2V4Real

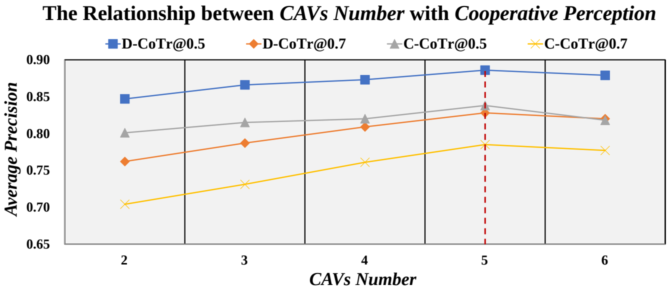


Fig. 5. Ablation study on the relationship between ***CAVs number*** with ***Cooperative Perception***. Notable, *D* and *C* denote the performance evaluated on OPV2V *Default* and *Culver City* splits under 0.5/0.7 thresholds, respectively.

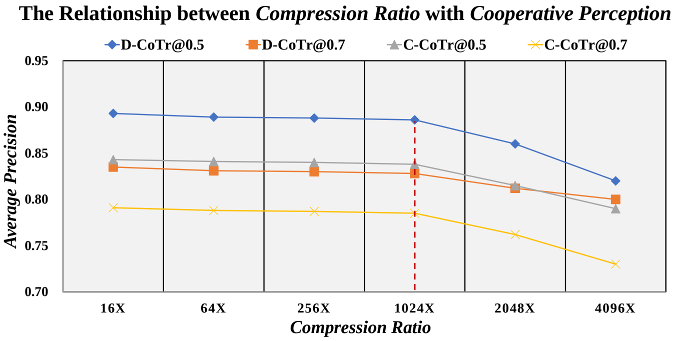


Fig. 6. Ablation study on the relationship between ***Compression Ratio*** with ***Cooperative Perception***. Notable, *D* and *C* denote the performance evaluated on OPV2V *Default* and *Culver City* splits under 0.5/0.7 thresholds, respectively.

over a substantial margin. Compared to CORE [52], it offers a 0.8%/1.8% AP boosts at both 0.5 and 0.7 levels on *Default* set, and simultaneously presents 1.6% AP@50 and 2.4% AP@70 increments on *Culver City* split. This drastic elevation suggests the advancement of our method. As for V2X-Sim, SECOND-*CoTr* still brings DiscoNet [50] by 1.2% (61.5% → 60.3%) and 1.1% (55.0% → 53.9%) precision under two thresholds, demonstrating its generalization across various datasets.

Furthermore, cooperative performance on V2V4Real *Test* is illustrated in Table IV, and we select PointPillars network with *CoTr* strategy for a fair comparison. Distinctly, PointPillars-*CoTr* achieves a AP of 65.4% and 36.1% at 0.5/0.7 thresholds on practical scenarios, respectively, which exceeds other single-agent and multi-agent counterparts by a considerable margin. Compared to V2VNet [14] and attentive fusion [23], it delivers 0.7% ∼ 0.9% and 1.8% ∼ 2.5% AP promotions in both IoUs, suggesting its competitiveness and advantage in physical environment. Nonetheless, our method falls behinds CoBEVT [59] with 1.1% AP@50 and V2X-ViT [51] with 0.8% AP@70. It is claimed that the real-world pattern is much challenging and confused, and a performant collaboration would be devoted to explore in the future.

Result Visualizations. Qualitative experiments are performed to evaluate collaborative perception performance intuitively, and we list visualization results of SECOND-*CoTr* on OPV2V and PointPillars-*CoTr* on V2V4Real splits for simplicity, are depicted in Fig.7∼Fig.8. Apparently, our proposed method showcases much powerful and robust to varying simulated/real scenarios: with the help of *CoTr* method, it could recognize hard samples in such occluded/crowded situation, largely enhancing the perception scope and probing range.

TABLE VI
ABLATION STUDY ON THE EFFECTIVENESS OF ***Anti-interference Ability*** ON ***OPV2V Default*** SPLIT. WE CONSIDER POSITION/HEADING ERROR AND TIME DELAY DISTURBANCES, AND ACCURACY DROP IS REPORTED IN THE BRACKET.

Methods		*OPV2V Default* → AP@50 (%)		
		OPV2V-attn	CoBEVT	PointPillars-*CoTr*
Position (*m*)	0.0	86.4	86.1	88.6
	0.1	84.9 (−1.5)	85.0 (−1.1)	88.2 (−0.4)
	0.2	82.1 (−2.8)	82.3 (−2.7)	86.5 (−1.7)
	0.3	78.5 (−3.6)	78.9 (−3.4)	84.3 (−2.2)
Heading (°)	0.0	86.4	86.1	88.6
	0.1	83.8 (−2.6)	84.2 (−1.9)	87.7 (−0.9)
	0.2	81.0 (−2.8)	82.7 (−1.5)	86.1 (−1.6)
	0.3	77.8 (−3.2)	80.6 (−2.1)	83.9 (−2.2)
Delay (*s*)	0.0	86.4	86.1	88.6
	0.1	84.7 (−1.7)	85.0 (−1.1)	87.9 (−0.7)
	0.2	83.0 (−1.7)	83.3 (−1.7)	86.3 (−1.6)
	0.3	79.9 (−3.1)	81.5 (−1.8)	84.9 (−1.4)

D. Ablation Study

We further conduct ablation analysis of collaborative strategy, CAV number, compression ratio, and anti-interference ability, to probe into their effectiveness in the context of multi-agent perception. For simplicity, all experiments adopt PointPillars trained on OPV2V as the baseline.

1) Effectiveness of Collaborative Strategy: table V lists the OPV2V *Default* result achieved by a variety of intermediate collaborations, *i.e.*, Average, Max [9], Mask-aware [14] and Attentive [23] fusions incorporated with PointPillars. Despite their promising progresses, our proposed *CoTr* leverages both spatial correlation and channel semantic fully, and clarifies the advantage of multi-agent collaboration with an evident AP raising of 20.7% and 22.6% at two thresholds.

2) Effectiveness of CAV number: fig.5 depicts the relationship between CAV number with cooperative perception on both OPV2V *Default* and *Culver City* sets. Evidently, detection accuracy tends to increase with the vehicle number linearly until 5 CAVs, conforming with the intuition of multi-view information for ego-vehicle perception enhancement. Whereas, performance drop occurs with more participants (*i.e.*, 6). We argue that five networked vehicles sufficiently provide 360° field-of-view surroundings for covering potential occlusion or blind-spot areas, and information redundancy would be detrimental to collaborative perception with the introduction of background noise or irrelevant object.

3) Effectiveness of Compression Ratio: the relationship between compression ratio and cooperative performance is investigated by changing convolution number in encoder to emulate the varying feature resolution during transmission. As illustrated in Fig.6, a new state-of-the-art cooperative detection performance is reported by PointPillars-*CoTr* without feature compression on both *Default* and *Culver City* splits. Intuitively, an acceptable decay emerges unavoidably with the decreasing feature resolution until $1024\times$ reduction. We speculate that key-point information loss would be caused by a larger downsampling process (*i.e.*, $2048\times$), and it is reasonable that $1024\times$ compression ratio could receive a satisfactory trade-off performance-efficiency result in terms of collaborative perception.

4) Effectiveness of Anti-interference Ability: the resistance of *CoTr* is also analyzed on OPV2V *Default* samples: Gaussian noise and uniform distribution are adopted for varying position/heading error and time delay simulation, and we explore the cooperative perception under different perturbations. Compared with two mainstream counterparts, *CoTr* manifests an advantageous robustness against real-world disturbances, and a considerable performance fall is reported as tabulated in Table VI. For instance, PointPillars-*CoTr* suffers from $0.4\% \sim 2.2\%$ accuracy degradation under a standard deviation of $[0.1, 0.3]$ position/heading error. In addition, it experiences $0.7/1.6/1.4$ AP declines when encountering different delay levels. We emphasize the stability and permanence of our proposed method, and more susceptibility analysis would be developed in the future.

V. Conclusion

In this paper, we develop *V2VFormer*, a novel collaborative perception framework with *Tr*ansformer-based *Co*llaboration (*CoTr*). Concretely, Spatial-Aware Transformer (*SAT*) is responsible for which agent should be collaborated according to spatial correlation among CAVs, while Channel-Wise Transformer (*CWT*) aims for sufficient semantic interaction across channels. Moreover, a new-built dataset V2V-Set is augmented on the top of OPV2V with a diversity of driving conditions. We conduct an extensive experiments on OPV2V, V2X-Sim and V2V4Real benchmarks, and our proposed method reports the state-of-the-art cooperative perception performance in both simulated/real scenarios, thereby demonstrating its superiority and advancement. Furthermore, abundant training samples would provide essential knowledge for performance promotion, and ablation study further reveals the effectiveness and robustness of each ingredient. We expect this work would shed a light on V2V perception in the future.

VI. Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under Grant No.2022YFC3803700, the National Natural Science Foundation of China under Grant No.U20A20155, Grant No.62173012, Grant No.52202391, and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant No.2023QNRC001.

References

[1] S. Chen, J. Hu, Y. Shi, L. Zhao, and W. Li, "A vision of c-v2x: Technologies, field testing, and challenges with chinese development," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 7, pp. 3872–3881, 2020.

[2] 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 Transactions on Intelligent Transportation Systems (TITS)*, 2022.

[3] Y. Cai, T. Luan, H. Gao, H. Wang, L. Chen, Y. Li, M. A. Sotelo, and Z. Li, "Yolov4-5d: An effective and efficient object detector for autonomous driving," *IEEE Transactions on Instrumentation and Measurement (TIM)*, vol. 70, p. 4503613, 2022.

[4] D. Tian, C. Lin, J. Zhou, X. Duan, Y. Cao, D. Zhao, and D. Cao, "Sa-yolov3: An efficient and accurate object detector using self-attention mechanism for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 23, pp. 4099–4110, 2022.

[5] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "3d-dfm: Anchor-free multimodal 3-d object detection with dynamic fusion module for autonomous driving," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2022.

[6] Z. Liu, Y. Cai, H. Wang, L. Chen, H. Gao, Y. Jia, and Y. Li, "Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 23, pp. 6640–6653, 2022.

[7] Y. Cai, L. Dai, H. Wang, L. Chen, and Y. Li, "Dlnet with training task conversion stream for precise semantic segmentation in actual traffic scene," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021.

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

[9] 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 *4th ACM/IEEE Symposium on Edge Computing (SEC)*, 2019, pp. 88–100.

[10] W. Arnold, O. Y. AI-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "Cooperative object classification for driving applications," in *IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 2484–2489.

[11] E. Arnold, M. Dianati, R. d. Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 23, pp. 1852–1864, 2020.

[12] J. B. Kenney, "Dedicated short-range communications (dsrc) standards in the united states," *Proceedings of the IEEE*, vol. 99, pp. 1162–1182, 2011.

[13] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 21 361–21 370.

[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 *European Conference on Computer Vision (ECCV)*, 2020, pp. 605–621.

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

[16] H. Caesar, A. H. Bankiti, Varun adn Land, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.

[17] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Gaine, and et al, "Scalability in perception for autonomous driving: Waymo open dataset," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2443–2451.

[18] C. Lin, D. Tian, X. Duan, and J. Zhou, "3d environmental perception modeling in the simulated autonomous-driving systems," *IEEE Complex System Modeling and Simulation (CSMS)*, vol. 1, pp. 45–54, 2021.

[19] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun, "Lidarsim: Realistic lidar simulation by leveraging the real world," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 167–11 176.

[20] 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 *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2021, pp. 1155–1162.

[21] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Annual Conference on Robot Learning CoRL*, 2017, pp. 1–16.

[22] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumosimulation of urban mobility: an overview," in *International Conference on Advances in System Simulation*, 2011, pp. 23–28.

[23] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2583–2589.

[24] H. Chen, B. Liu, X. Zhang, F. Qian, Z. Mao, and F. Yiheng, "A cooperative perception environment for traffic operations and control," *arXiv arXiv prePrint: 2208.02792*, 2022.

[25] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, H. Yu, B. Zhou, and J. Ma, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13 712–13 722.

[26] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4490–4499.

[27] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 1951–1960.

[28] Q. Jiang, C. Hu, B. Zhao, Y. Huang, and X. Zhang, "Scalable 3d object detection pipeline with center-based sequential feature aggregation for intelligent vehicles," *IEEE Transactions on Intelligent Vehicles (TIV)*, pp. 1–12, 2023.

[29] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[30] C. He, H. Zeng, J. Huang, X. Hua, and L. Zhang, "Structure aware single-stage 3d object detection from point cloud," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 873–11 882.

[31] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel r-cnn: Towards high performance voxel-based 3d object detection," in *The AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[32] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast point r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 9775–9784.

[33] J. Noh, S. Lee, and B. Han, "Hvpr: Hybrid voxel-point representation for single-stage 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 605–14 614.

[34] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.

[35] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 770–779.

[36] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 040–11 048.

[37] C. Chen, Z. Chen, J. Zhang, and D. Tao, "Sasa: Semantics-augmented set abstraction for point-based 3d object detection," in *The AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[38] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensor*, vol. 18, no. 10, p. 3337, 2018.

[39] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 698–12 705.

[40] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[41] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X. Hua, and M. Zhao, "Improving 3d object detection with channel-wise transformer," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 2743–2752.

[42] C. He, R. Li, S. Li, and L. Zhang, "Voxel set transformer: A set-to-set approach to 3d object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8417–8427.

[43] X. Pan, Z. Xia, S. Song, L. E. Li, and G. Huang, "3d object detection with pointformer," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7463–7472.

[44] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 3164–3173.

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

[46] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4106–4115.

[47] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "Hydro-3d: Hybrid object detection and tracking for cooperative perception using 3d lidar," *IEEE Transactions on Intelligent Vehicles (T-IV)*, early access 2023.

[48] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles (TIV)*, vol. 8, pp. 2650–2660, 2023.

[49] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 6035–6042.

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

[51] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 107–124.

[52] B. Wang, L. Zhang, Z. Wang, Y. Zhao, and T. Zhou, "Core: Cooperative reconstruction for multi-agent perception," *arXiv arXiv prePrint: 2307.11514*, 2023.

[53] S.-W. Kim, B. Qin, Z. Chong, X. Shen, W. Liu, E. Frazzoli, and D. Rus, "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Transactions on Intelligent Transportation Systems (TITS)*, vol. 16, pp. 663–680, 2015.

[54] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 953–959.

[55] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 1616 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021, pp. 1–21.

[56] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[57] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6881–6890.

[58] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y.-W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International Conference on Machine Learning (ICML)*, 2019, pp. 3744–3753.

[59] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative birds eye view semantic segmentation with sparse transformers," in *Annual Conference on Robot Learning (CoRL)*, 2022, pp. 1–12.

[60] T.-y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[61] Y. Yuan, H. Cheng, and S. Monika, "Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving," *IEEE Robotics and Automation Letter (RA-L)*, vol. 7, no. 2, pp. 3054–3061, 2022.

[62] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 4874–4886.

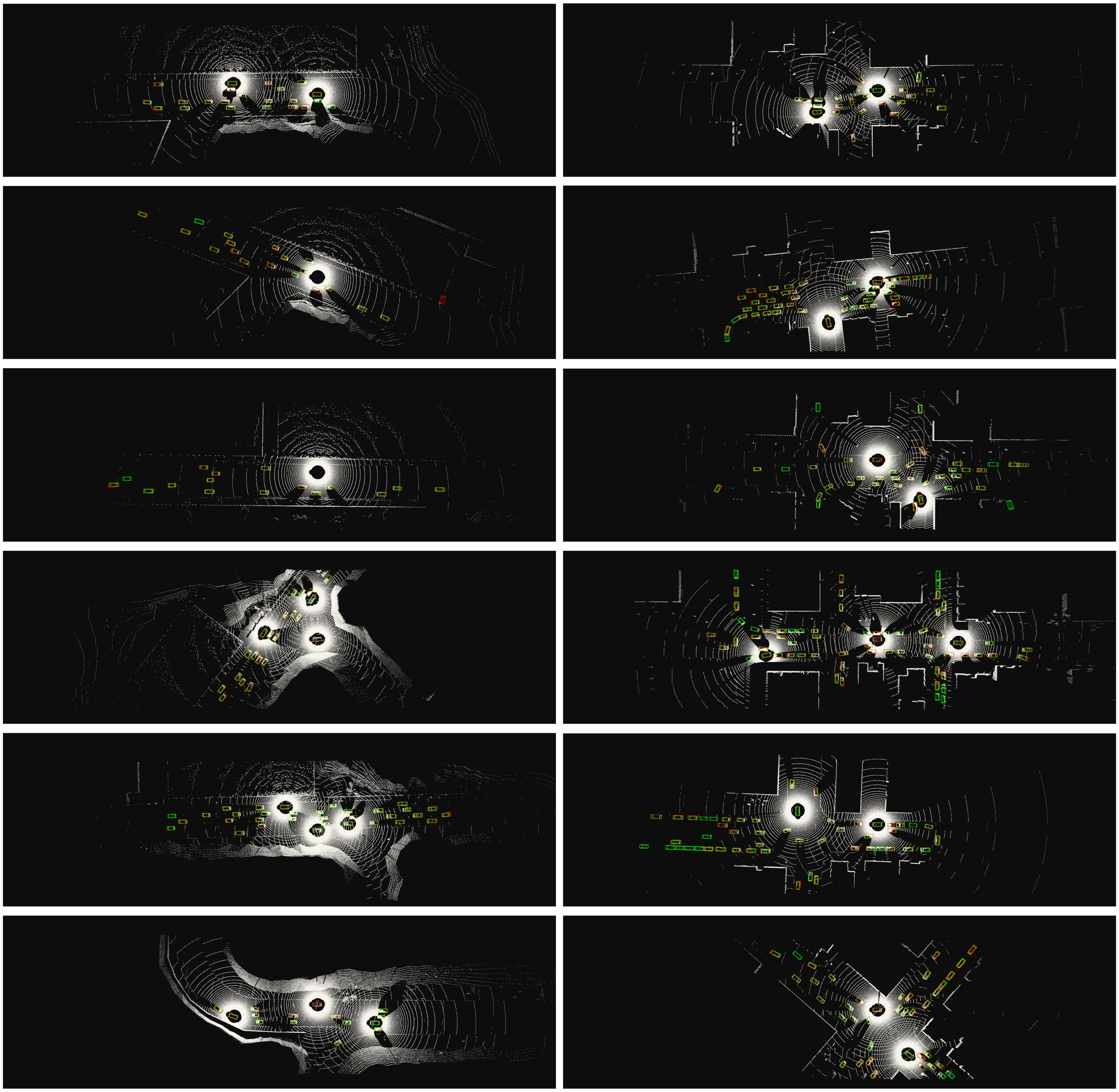


Fig. 7. Visualization results achieved by **SECOND-*CoTr*** on **OPV2V** ***Default*** and ***Culver City*** splits at the *left* and *right* columns, respectively. It covers a variety of simulation condition, and we draw ground-truth (GT) and prediction with *Green* and *Red* rectangles. Best Viewed in color.

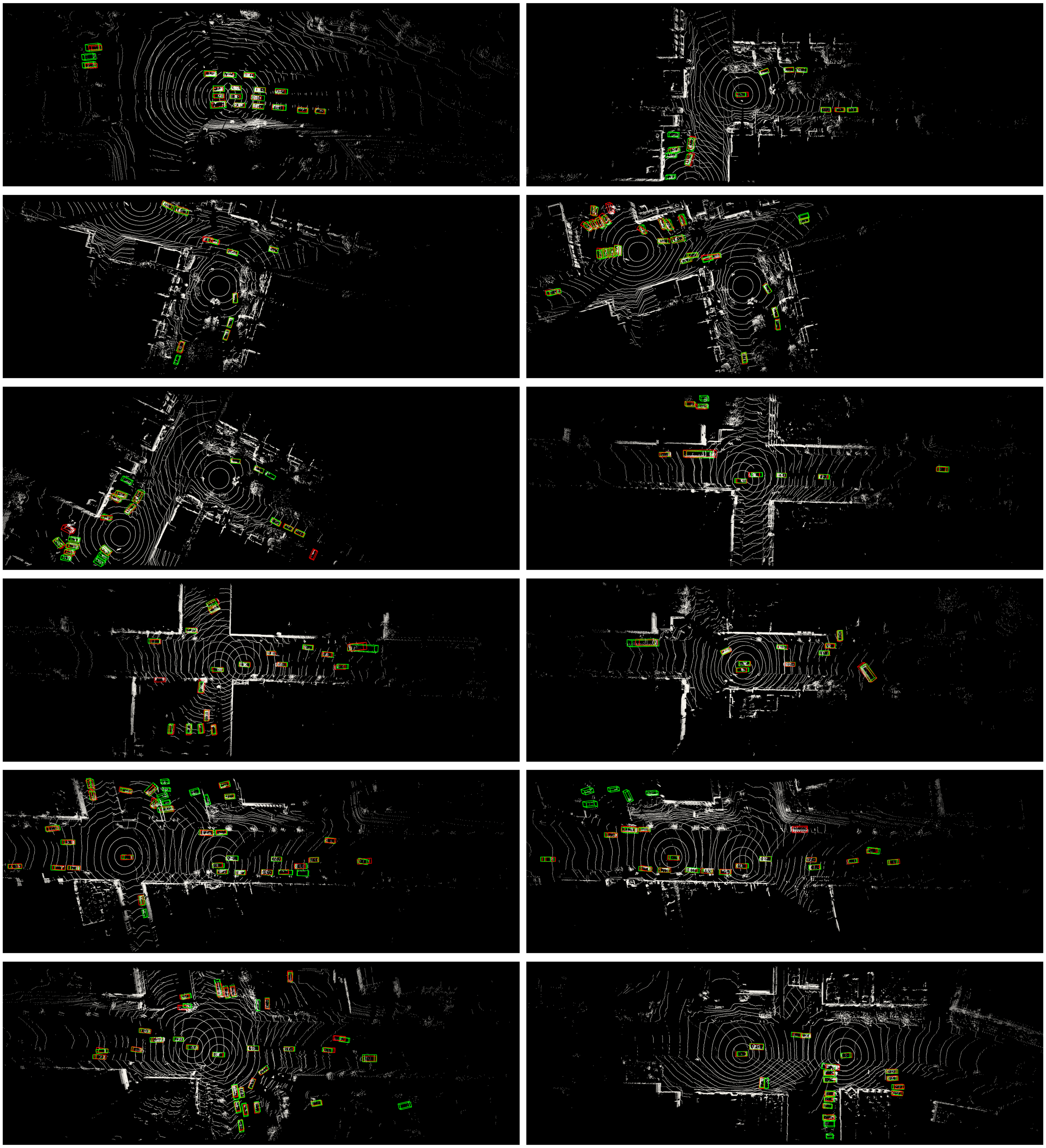


Fig. 8. Visualization results achieved by **PointPillars-*CoTr*** on **V2V4Real *Test*** split covered a diversity of real-world scenarios. The ground-truth (GT) and prediction are drawn with *Green* and *Red*, respectively. Best Viewed in color.

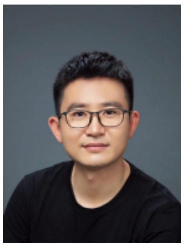
Chunmian Lin received the Ph.D degree in Electronic and Information from Beihang University, Beijing, China. He is currently a post-doctoral fellow in the School of Transportation Science and Engineering, Beihang University, Beijing, China. His research interests include but not limited in autonomous driving, image processing, computer vision, information fusion and deep learning, particularly their applications in intelligent transportation systems.

Daxin Tian [M'13, SM'16] is currently a professor in the School of Transportation Science and Engineering, Beihang University, Beijing, China. He is IEEE Senior Member, IEEE Intelligent Transportation Systems Society Member, and IEEE Vehicular Technology Society Member, etc. His current research interests include mobile computing, intelligent transportation systems, vehicular ad hoc networks, and swarm intelligent.

Xuting Duan 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 system and internet of vehicles.

Jianshan Zhou received the B.Sc. and M.Sc. degrees in Traffic Information Engineering and Control from Beihang University in 2013 and 2016, respectively. He is currently working towards the Ph.D. degree with the School of Transportation Science and Engineering, Beihang University, Beijing, China. His current research interests include wireless communication, artificial intelligent system, and intelligent transportation systems.

Dezong Zhao [M'12, SM'17] 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. Dr Zhaos research interests include connected and autonomous vehicles, machine learning and control engineering. His work has been recognised 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 a Professor at Tsinghua University. His current research interests include driver cognition, automated driving and cognitive autonomous driving. He has contributed more than 200 papers and 3 books. He received the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award and 6 Best Paper Awards from international conferences. Prof. Cao has served as Deputy Editor-in-Chief for IET Intelligent Transport Systems Journal and Associate Editors for IEEE Transactions on Vehicular Technology, IEEE Transactions on Intelligent Transportation Systems, IEEE Transactions on Intelligent Vehicles, IEEE/ASME Transactions on Mechatronics, IEEE Transactions on Industrial Electronics, IEEE/CAA Journal of Automatica Sinica, IEEE Transactions on Computational Social Systems, and ASME Journal of Dynamic Systems, Measurement, and Control. Prof. Cao is an IEEE VTS Distinguished Lecturer.