Learning for Vehicle-to-Vehicle Cooperative Perception Under Lossy Communication

Jinlong Li, *Graduate Student Member, IEEE*, Runsheng Xu, Xinyu Liu, *Graduate Student Member, IEEE*, Jin Ma, Zicheng Chi, *Member, IEEE*, Jiaqi Ma, *Member, IEEE*, and Hongkai Yu, *Member, IEEE*

***Abstract*—Deep learning has been widely used in intelligent vehicle driving perception systems, such as 3D object detection. One promising technique is Cooperative Perception, which leverages Vehicle-to-Vehicle (V2V) communication to share deep learning-based features among vehicles. However, most cooperative perception algorithms assume ideal communication and do not consider the impact of Lossy Communication (LC), which is very common in the real world, on feature sharing. In this paper, we explore the effects of LC on Cooperative Perception and propose a novel approach to mitigate these effects. Our approach includes an LC-aware Repair Network (LCRN) and a V2V Attention Module (V2VAM) with intra-vehicle attention and uncertainty-aware inter-vehicle attention. We demonstrate the effectiveness of our approach on the public OPV2V dataset (a digital-twin simulated dataset) using point cloud-based 3D object detection. Our results show that our approach improves detection performance under lossy V2V communication. Specifically, our proposed method achieves a significant improvement in Average Precision compared to the state-of-the-art cooperative perception algorithms, which proves the capability of our approach to effectively mitigate the negative impact of LC and enhance the interaction between the ego vehicle and other vehicles.**

***Index Terms*—Deep learning, vehicle-to-vehicle cooperative perception, 3D object detection, lossy communication, digital twin.**

I. Introduction

How to perceive the surrounding objects precisely in complex real-world scenarios is critical for modern intelligent vehicle research. The accurate perception system (e.g., 3D object detection) is the fundamental base for the next motion planning and control of the intelligent vehicles, which implies tremendous impacts on the driving safety of intelligent vehicles [1], [2], [3], [4], [5], [6].

Because of the perception limitation of the current individual intelligent vehicle [7], [8], [9], the cooperative perception of Connected Automated Vehicles (CAV) recently attracted much attention in this research community. Compared to the perception of individual intelligent vehicles, recent studies [10], [11], [12] show that cooperative perception of CAV can significantly improve the perception performance by leveraging Vehicle-to-Vehicle (V2V) communication technology for information sharing. Information sharing through V2V communication is an important technology for CAV cooperative perception, which is utilized to observe a wider range and perceive more occluded objects in the complex traffic environment [13], [14]. There are three ways for information sharing during V2V communication: (1) sharing raw sensor data as early fusion, (2) sharing intermediate features of the deep learning based detection networks as intermediate fusion, (3) sharing detection results as late fusion. Recent state-of-the-art studies [10], [15] show that intermediate fusion is the best trade-off between detection accuracy and bandwidth requirement. This paper also focuses on the intermediate fusion during communication for V2V cooperative perception.

Many intermediate fusion methods have been recently proposed for the V2V cooperative perception [10], [11], [12], [13]; however, all of them assume the ideal communication. The only V2V cooperative perception study that considered non-ideal communication focused solely on communication delays [15]. To date, no existing work has explored the impact of Lossy Communication (LC) on V2V cooperative perception in complex real-world driving environments. In urban traffic scenarios, V2V communication is susceptible to a range of factors that can result in lossy communication, such as multi-path effects from obstacles (e.g., buildings and vehicles) [17], Doppler shift introduced by fast-moving vehicles [18], interference generated by other communication networks [19], and dynamic topology caused by routing failures [20], as well as various weather conditions. Incomplete or inaccurate shared intermediate features resulting from lossy communication could compromise the effectiveness and efficiency of V2V cooperative perception, as shown in Fig. 1. Failure to address LC in cooperative perception could lead to degraded perception performance, increased collision risk, and reduced traffic efficiency.

This paper first studies the negative effect of lossy communication in the V2V cooperative perception and then proposes a novel intermediate LC-aware feature fusion method to address the issue. Specifically, the proposed method includes an LC-aware Repair Network (LCRN) to recover the incomplete shared features by lossy communication and a specially

Manuscript received 28 February 2023; accepted 7 March 2023. Date of publication 21 March 2023; date of current version 19 May 2023. This work was supported in part by NSF under Grants 2215388 and CNS-2127881 and in part by OpenCDA Ecosystem. *(Corresponding author: Hongkai Yu.)*

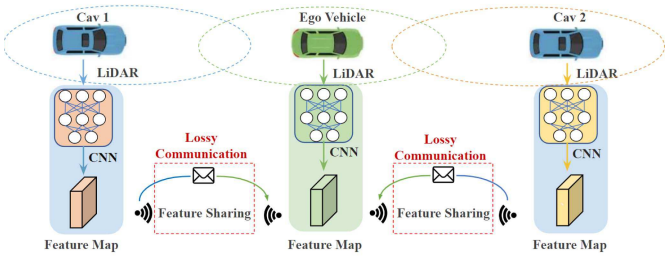
Jinlong Li, Xinyu Liu, Jin Ma, Zicheng Chi, and Hongkai Yu are with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH 44115 USA (e-mail: lijinlong1117@gmail.com; z2xinyu16@gmail.com; majinwakeup@gmail.com; z.chi@csuohio.edu; h.yu19@csuohio.edu).

Runsheng Xu and Jiaqi Ma are with the Department of Civil and Environmental Engineering, University of California, Los Angeles, CA 90024 USA (e-mail: rxx3386@g.ucla.edu; jiaqima@ucla.edu).

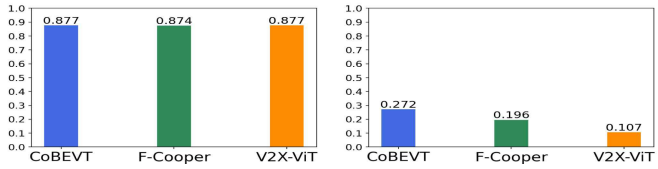
The code is available at: https://github.com/jinlong17/V2VLC.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2023.3260040.

Digital Object Identifier 10.1109/TIV.2023.3260040



(a) Pipeline of V2V Cooperative Perception with lossy communication



(b) Detection in ideal communication (c) Detection in lossy communication

Fig. 1. Illustration of the V2V cooperative perception pipeline and its detection performance drop suffering from lossy communication on the public digital-twin CARLA simulator [16] based OPV2V dataset [10], where three intermediate fusion methods all trained under ideal communication are displayed: CoBEVT [11], F-Cooper [12], and V2X-ViT [15].

designed V2V Attention Module (V2VAM) to enhance the interaction between the ego vehicle and other vehicles. The V2VAM includes the intra-vehicle attention of ego vehicle and uncertainty-aware inter-vehicle attention. It is challenging to collect the authentic CAV perception data with lossy communication in real-world driving, and considering the advantage of the digital twin in many application [21], [22], [23], [24], [25], [26], this paper evaluates the proposed method in a digital-twin CARLA simulator [16] based public cooperative perception dataset OPV2V [10]. The contributions of this paper are summarized as follows.

- We propose the first research on V2V cooperative perception (point cloud-based 3D object detection) under lossy communication and study the side effect of lossy communication on cooperative perception, specifically the impact on detection performance.
- This paper proposes a novel intermediate LC-aware feature fusion method to relieve the side effect of lossy communication by a LC-aware Repair Network and enhance the interaction between the ego vehicle and other vehicles by a specially designed V2V Attention Module including intra-vehicle attention of ego vehicle and uncertainty-aware inter-vehicle attention.
- We evaluate the proposed method on the public cooperative perception dataset OPV2V, which is based on the digital-twin CARLA simulator [16].

The rest of this paper is organized as follows. Section II briefly reviews the related literature to this work, Section III presents the details of the proposed V2V cooperative perception method under lossy communication, Section IV provides the experiments and analysis with two scenarios: Ideal Communication and Lossy Communication, and the final conclusion are given in Section V.

II. Related Work

A. 3D Perception for Autonomous Driving

3D object detection is one of the most critical ways to the success of autonomous driving perception. Based on recently available sensor modality [27], 3D detection method can be roughly divided into three categories: (1) **Camera-based detection methods** where approaches detect 3D objects using a single or multiple RGB images [28], [29], [30]. For example, CaDDN [28] utilizes the depth distributions combined with the image features to generate bird's-eye-view representations for 3D object detection. ImVoxelNet [29] constructs a 3D volume in 3D space and samples multi-view features to obtain the voxel representation for 3D object detection. DETR3D [30] uses queries to index into extracted 2D multi-camera features to directly estimate 3D bounding boxes in 3D spaces. (2) **LiDAR-based detection methods** where these methods typically convert LiDAR points into voxels or pillars, leading in voxel-based [31], [32] or pillar-based object detection methods [33], [34], [35]. PointRCNN [36] proposes a two-stage strategy based on raw point clouds, which learns rough estimation first and then refines it with semantic attributes. Some methods [31], [32] propose to split the space into voxels and produce features per voxel. However, 3D voxels are usually expensive to process. To address this issue, PointPillars [37] propose to compress all the voxels along the z-axis into a single pillar, then predict 3D boxes in the bird's-eye-view space. Moreover, some recent methods [38], [39] combine voxel-based and point-based approaches to detect 3D objects jointly. (3) **Camera-LiDAR fusion detection method** where it presents an approach fusing information from both image and LiDAR points, which is a trend in 3D object detection. How to align the image features with point clouds is challenging in multimodal fusion. To solve this challenge, some methods [40], [41], [42] use a two-step framework, where detecting the object in 2D images in the first stage, then using the obtained information to further process point clouds for 3D detection. While other works [43], [44] develop end-to-end fusion pipelines and leverage cross-attention mechanisms to perform feature alignment. Our work in this paper focuses on the cooperative point cloud based 3D object detection to achieve fast processing and real-time performance [33], [34]; pillar-based approach would be used in our following experiments.

B. Vehicle-to-Vehicle Cooperative Perception

The performance of a 3D perception method highly depends on the accuracy of 3D point clouds. However, LiDAR cameras suffer from refraction, occlusion, and long-range distance, so the single-vehicle system could become unreliable under some challenging situations [10]. In recent years, Vehicle-to-Vehicle (V2V) / vehicle-to-infrastructure (V2X) cooperating system has been proposed to overcome the disadvantages of the single-vehicle system by using multiple vehicles. The collaboration among different vehicles enables the 3D perception network to fuse information from different sources.

Some former methods use **Early fusion** to share raw data among different vehicles. For example, Cooper [45] fused the

point clouds from different connected autonomous vehicles and made predictions based on the aligned data. AUTOCAST [46] exchanged sensor readings from different sensors to broaden the perceptive field for a single vehicle. Other methods use **Late fusion** to integrate the 3D detection results from each vehicle. Rawashdeh et al. [47] proposed a machine learning based method that shares the dimension and the location of the center point for each tracked object. Some other late fusion methods [48], [49] also adopt point clouds as sensor data from both vehicle and infrastructure. While early fusion requires large bandwidth and data transfer speed, late fusion may generate undesirable results due to the biased individual prediction. In order to find the balance between data load and accuracy, recent methods focus on **Intermediate fusion** by sharing intermediate representations. F-cooper [12] applied voxel features fusion and spatial feature fusion from two cars. V2VNet [13] employed a graph neural network to aggregate features extracted by LiDAR from each vehicle. V2X-ViT [15] proposed a vision Transformer architecture to fuse features from vehicles and infrastructures. Cui et al. [50] proposed a Point-based Transformer for point cloud processing, which can integrate collaborative perception with control decisions. Tu et al. [51] proposed an efficient and practical online attack network in a multi-agent deep learning system based on intermediate representations. Luo et al. [52] utilized attention modules to fuse the intermediate feature and enhance feature complementarity. Lei et al. [53] proposed a latency compensation module to realize intermediate feature-level synchronization. Hu et al. [54] proposed a spatial confidence-aware communication strategy to use less communication to improve performance by focusing on perceptually critical areas. OPV2V [10] utilized a self-attention module to fuse the received intermediate features. CoBEVT [11] proposed local-global sparse attention that captures complex spatial interactions across views and agents to improve the performance of cooperative perception. However, these fusion methods are all with the assumption of ideal communication, which would suffer dramatic performance drop with lossy communication in the real world. To address this issue, we design a special V2V Attention Module (V2VAM), including intra-vehicle attention of ego vehicle and uncertainty-aware inter-vehicle attention to enhance the V2V interaction.

C. Communication Issue in V2V Perception

V2V and V2X communication can improve the safety and reliability of autonomous vehicles by exchanging information with surrounding vehicles. However, communication among vehicles may bring new issues to this research field [55]. Due to the nature of the connectivity, lossy communication is inevitable in wireless channels. Some factors like channel errors, network congestion, and delay deadline violation can cause packet losses during the transmission of data in the wireless network [56]. Low latency and high reliability are two common challenges for V2V communication. For example, in the pre-crash sensing scene, the maximum latency is only 20 ms, and data delivery reliability must be greater than 99% [55], [57]. Several works have proposed specific resource allocation schemes to ensure latency and reliability of V2V communication systems by using Lagrange dual decomposition and binary search [58], greedy link selection [59], or federated learning [60]. Some studies also aim to improve the V2V communication security from different aspects such as authentication, data integrity, and data protection [61].

Lossy Communication (LC) is also a critical issue in V2V communication. According to studies on single-hop broadcasting, the obstacle (vehicles, buildings, etc.) between transmitter and receiver will result in signal power fluctuations, thus causing packet loss [21], [55], [62], [63]. The shared data could also be interfered with by other signals or modified by attackers before arriving at its destination, thus leading to lossy data. In this work, we aim to eliminate the lossy communication by proposing an LC-aware repair network and improving the robustness of the V2V perception network.

III. Methodology

In this paper, we focus on the cooperative LIDAR-based 3D object detection task for autonomous driving and consider a realistic scenario where communication loss exists in collaboration. Since we focus on the lossy communication challenge during data transmission in this paper, we assume there are no communication delays or localization errors in the V2V system. To handle lossy communication challenges in the real world and enhance CAV's cooperative perception capability, inspired by [10], this paper proposes a novel intermediate LC-aware feature fusion framework. The overall architecture of the proposed framework is illustrated in Fig. 2, which includes five components: 1) V2V metadata sharing, 2) LIDAR feature extraction, 3) Feature sharing, 4) LC-aware repair network and V2V Attention module, 5) classification and regression headers.

A. Overview of Architecture

V2V metadata sharing: We select one of the CAVs as the ego vehicle to construct a spatial graph around it where each node is a CAV within the communication range, and each edge represents a directional V2V communication channel between a pair of nodes. Upon receiving the relative pose and extrinsic of the ego vehicle, all the other CAVs nearby will project their own LiDAR point clouds to the ego vehicle's coordinate frame before feature extraction, which could be simply formulated as

$$p^t_{cav_{projected}} = T_{cav\to ego} \cdot p^t_{cav}, \tag{1}$$

where p^t_{cav} is the CAV pose $[x, y, z, 1]^{\mathrm{T}}$ in i-th CAV at the time t, and $T_{cav\to ego} \in \mathbb{R}^{4\times 4}$ is coordinate transformation matrix from CAV to ego.

LIDAR feature extraction: The anchor-based PointPillar method [37] is selected as the 3D detection backbone to extract visual features from point clouds. Since it can be deployed in the real world easily than other 3D detection backbones (e.g. SECOND [31], PIXOR [64], and VoxelNet [32]) thanks to its low inference latency and optimized memory usage [10]. This method converts the raw point clouds to a stacked pillar tensor, then scattered to a 2D pseudo-image and fed to the PointPillar backbone. Finally, the backbone extracts informative visual

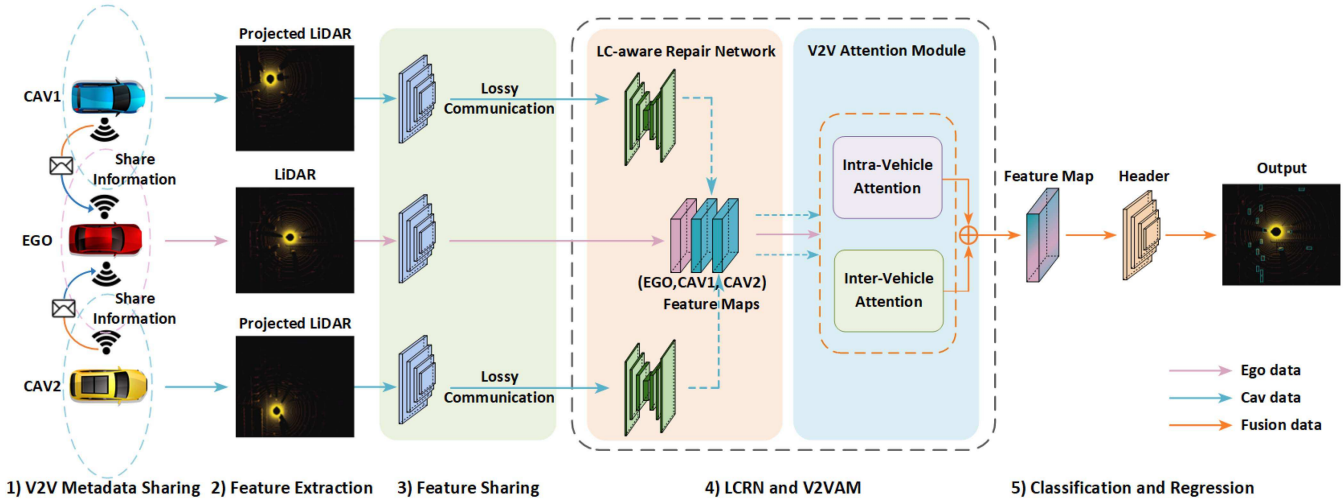


Fig. 2. The architecture of LC-aware feature fusion framework. The proposed model includes five components: 1) V2V Metadata Sharing, 2) LiDAR Feature Extraction, 3) Feature Sharing, 4) LC-aware Repair Network (LCRN) and V2V Attention Module (V2VAM), 5) Classification and Regression header.

TABLE I
3D OBJECT DETECTION PERFORMANCE COMPARISON ON TWO TESTING SETS OF OPV2V BASED THE TRAINING OF *SCHEME I*. WE SHOW AVERAGE PRECISION (AP) AT IOU=0.5, 0.7. NOTE THAT V2VAM IS ONLY OUR PROPOSED V2V ATTENTION MODULE WHILE V2VAM+LCRN IS OUR FULL PROPOSED METHOD

Method	Com. Type	V2V CARLA Towns AP@ 0.5	V2V CARLA Towns AP@ 0.7	V2V Culver City AP@ 0.5	V2V Culver City AP@ 0.7
NO Fusion	Ideal	0.679	0.602	0.557	0.471
	Lossy	0.679	0.602	0.557	0.471
F-Cooper [12]	Ideal	0.844	0.743	0.874	0.715
	Lossy	0.036	0.029	0.196	0.146
V2VNet [13]	Ideal	0.874	0.712	0.855	0.630
	Lossy	0.024	0.014	0.102	0.061
OPV2V [10]	Ideal	0.871	0.793	0.868	0.745
	Lossy	0.015	0.011	0.010	0.006
CoBEVT [11]	Ideal	0.914	0.836	0.877	0.748
	Lossy	0.089	0.069	0.272	0.202
V2X-ViT [15]	Ideal	0.840	0.726	0.877	0.720
	Lossy	0.083	0.054	0.107	0.067
V2VAM	Ideal	**0.926**	**0.861**	**0.885**	**0.785**
	Lossy	0.085	0.075	0.095	0.070

feature maps. Each CAV has its own LIDAR feature extraction module.

Feature sharing: In this component, the ego vehicle will receive the neighboring CAV feature maps after each CAV feature extraction, and these received intermediate features will be fed into the remaining detection networks in the ego vehicle. In the real-world scenario (e.g. urban building and unpredictable occlusion), the transmission of the feature maps usually suffers inevitable damage that leads to lossy communication. As a result, existing 3D object detectors would suffer a dramatic performance drop with the lossy features collected from surrounding CAVs, as shown in Table I.

LC-aware Repair Network and V2V Attention Module: The intermediate features aggregated from other surrounding CAVs are fed into the major component of our framework *i.e.*, LC-Aware Repair Network for recovering the intermediate feature map in lossy communication by using tensor-wise filtering, and V2V Attention module for iterative inter-vehicle as well as intra-vehicle feature fusion utilizing attention mechanisms. The proposed LC-aware repair network and V2V attention module will be revealed with details in Section III-B and Section III-C, respectively.

Classification and regression headers: After receiving the final fused feature maps, two prediction headers are utilized for box regression and classification.

B. LC-Aware Repair Network

Image denoising is one of the longstanding challenging tasks in computer vision. The primary sources of noise [65] are shot noise, where a Poisson process with variance equal to the signal level, and read noise, where an approximately Gaussian process is caused by diverse sensor readout effects. To denoise them, some deep learning-based methods [66], [67], [68] use denoising networks that generate a filter for every pixel in the desired output to constrain the output space and thereby prevent the impact of artifacts. Inspired by these architectures, to handle the common V2V communication challenges *i.e.*, lossy communication, we design a customized LC-aware repair network for intermediate feature recovering from other CAVs.

The framework of the LC-aware repair network is shown in Fig. 3, which is an encoder-decoder architecture with skip connections. This network generates a specific per-tensor filter kernel to jointly align and recover the input damaged feature to produce a recovered version of the output feature. The input feature for LC-aware repair network is $S \in \mathbb{R}^{c \times h \times w}$, then a tensor-wise kernel K is generated and applied to S to produce the recovered output feature $\hat{S} \in \mathbb{R}^{c \times h \times w}$. the specific tensor-wise filter kernel could be simply formulated as

$$K = Conv(S), \tag{2}$$

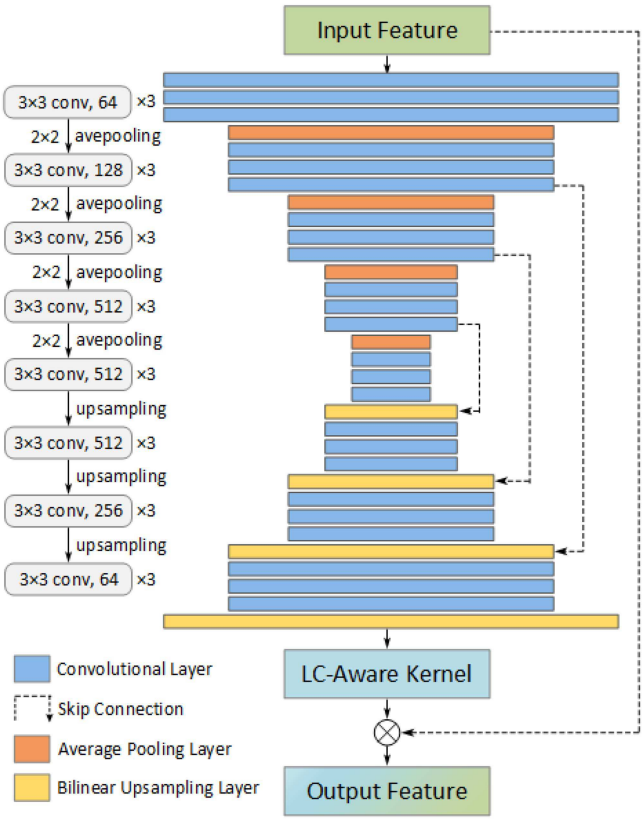


Fig. 3. Illustration of LC-aware Repair Network. The LC-aware Repair architecture for feature recovery is based on the encoder-decoder structure, which outputs per-tensor feature kernels. These kernels then are applied to the input lossy features.

and the value at each tensor t in our output feature $\hat{S}$ is

$$\hat{S}^t = K^t \circledast S^t, \tag{3}$$

where $\circledast$ denotes the matrix dot product. $K \in \mathbb{R}^{(k\times k)\times h\times w}$ is a tensor-wise kernel, and each tensor in channel dimension $K^t \in \mathbb{R}^{k\times k}$ is a per-tensor kernel and can be applied to the $k \times k$ neighborhood region of each tensor t in the input feature $S \in \mathbb{R}^{c\times h\times w}$by multiplication. The $Conv(.)$ denotes the tensor-wise network and is used to perceive the input feature and generate the suitable kernel for each tensor.

To acquire the repaired output feature $\hat{S}$, the tensor-wise filtering $\circledast$ of the input damaged feature could largely preserve the feature detailed without corruption. Therefore, a large kernel size k is desired to leverage the rich neighborhood information of each tensor fully. In our experiment, the kernel size k is set to 5 due to memory limitations.

The LC-aware repair loss function $\mathcal{L}_{LC}(\hat{S}, \hat{S}^g)$ is the tensor-wise $L1$ distance between the ground truth original feature $\hat{S}^g$ before suffering lossy communication and the repaired feature $\hat{S}$. The repair loss can be defined as

$$\mathcal{L}_{LC}(\hat{S}, \hat{S}^g) = \|\hat{S}^g - \hat{S}\|. \tag{4}$$

C. V2V Attention Module

Self-attention mechanism [69] has emerged as a recent advance to capture long-range interaction; The key idea of self-attention is to calculate the response at a position as a weighted sum of the features at all locations, with the interaction between features determined by the features themselves rather than their relative location, as in convolutions. In this paper, after receiving the recovered intermediate feature, we aim to leverage the intermediate deep learning features from multiple nearby CAVs to improve perception performance based on V2V communication. We design a customized intra-vehicle and inter-vehicle attention fusion method by considering the lossy communication situation to enhance interaction between ego CAV and other CAVs. Moreover, we adopt a criss-cross attention module in our proposed V2V attention method, which can be leveraged to capture contextual information from full-feature dependencies more efficiently and effectively.

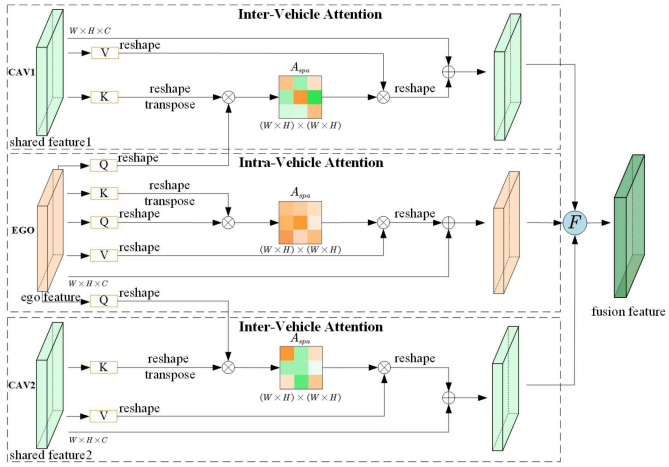


Fig. 4. The architecture of V2V Attention Module includes the intra-vehicle attention of ego vehicle and uncertainty-aware inter-vehicle attention. The final output is a fusion feature with interaction between ego features and other shared features from other CAVs. $\boldsymbol{F}$ denotes a set of max pooling, average pooling, and convolution operation.

Intra-Vehicle Attention: For the ego vehicle only, the intra-vehicle attention module can enable features from any position to perceive globally, thus enjoying full-image contextual information to better capture the representative feature. Formally, let $\mathbf{H}^e \in \mathbb{R}^{C\times H\times W}$ be an input feature map of an ego vehicle, which is perfect data generated by self-vehicle without suffering any lossy communication. In the intra-vehicle attention module, the feature map $\mathbf{H}^e$ would be calculated by three 1×1 convolutional layers to produce three feature vectors $\mathbf{Q}^e$, $\mathbf{K}^e$, and $\mathbf{V}^e$, respectively, where all of them have the same size, $\{\mathbf{Q}^e, \mathbf{K}^e, \mathbf{V}^e\} \in \mathbb{R}^{C\times H\times W}$. Following the scaled dot-product attention in [69], we compute the dot products of the $\mathbf{Q}^e$ and $\mathbf{K}^e$, then divide them using a scaling factor *i.e.* dimension of feature vectors, and apply a softmax function to obtain the weights on the $\mathbf{V}^e$. The intra-vehicle attention as shown in Fig. 4 is defined as follows,

$$\mathbf{A}^{intra} = \text{softmax}\left(\frac{\mathbf{Q}^e\mathbf{K}^{e\mathrm{T}}}{\sqrt{d_k^e}}\right)\mathbf{V}^e, \tag{5}$$

where d_k^e is the dimension of $\mathbf{K}^e$, and the standard $\text{softmax}()$ function is used as the activated function here. $\mathbf{A}^{intra}$ denotes

the output feature map of ego vehicle with considering all spatial information of the feature map.

Uncertainty-Aware Inter-Vehicle Attention: In V2V cooperative perception, the intermediate feature maps $\mathbf{H}^s \in \mathbb{R}^{C\times H\times W}$ aggregated from other CAVs are shared to the ego vehicle. The shared feature maps $\mathbf{H}^s$ with lossy communication would be recovered by LC-aware repair network, as introduced in Section III-B, but they are still noisy to some extent, while the ego feature maps $\mathbf{H}^e$ are prefect without any lossy transmission. Fusing these uncertain feature maps with a certain ego feature map directly could be risky in the cooperative perception interaction process. To address this issue, we propose an uncertainty-aware inter-vehicle attention fusion method by considering the uncertainty of the recovered feature maps. In this module, the shared feature maps would be calculated by two 1×1 convolutional layers to produce two feature vectors $\mathbf{K}^s$, and $\mathbf{V}^s$, respectively, where all of them have the identical size, $\{\mathbf{K}^s, \mathbf{V}^s\} \in \mathbb{R}^{C\times H\times W}$ and the other feature vector $\mathbf{Q}^e$ is directly obtained from ego self-vehicle instead of other vehicles, as shown in Fig. 4. Similar to the intra-vehicle attention in Section III-C, the uncertainty-aware inter-vehicle attention can be defined as

$$\mathbf{A}^{inter} = \sum_{i}^{N} \text{softmax}\left(\frac{\mathbf{Q}^e \mathbf{K}_i^{s\mathrm{T}}}{\sqrt{d_k^s}}\right)\mathbf{V}_i^s, \tag{6}$$

where d_k is the dimension of $\mathbf{K}_i^s$, and N is the number of the neighboring CAVs. $\mathbf{A}^{inter}$ denotes the sum of the output feature map considering the interaction between the ego vehicle and other vehicles.

Efficient Implementation: Inspired by [70], we adopt two consecutive criss-cross (CC) attention modules to implement V2V attention in point cloud data rather than scaled dot-product attention. The latter generates huge attention maps to measure the relationships for each point-pair, resulting in a very high complexity of $O((H \times W)^2)$, where $H \times W$ is the size of input features $\mathbf{H}^e$ and $\mathbf{H}^s$. The CC attention module [70] aggregates contextual information in horizontal and vertical directions, collecting contextual information from all pixels by serially stacking two CC attention modules. Each position has sparse connections to other positions in the feature map, with a total of $(H + W - 1)$ connections per position. This approach greatly reduces the complexity from $O((H \times W) \times (H \times W))$ to $O((H \times W) \times (H + W - 1))$ while still effectively capturing the relevant context from all vehicles through V2V communication.

After obtaining the intra-vehicle attention and inter-vehicle attention, all of them would be fed into the max pooling and average pooling layers separately to obtain abundant spatial information, then they are concatenated as the input for the 2D convolutional layer with ReLU activation function. Therefore, the final fusion feature output $\mathbf{A}^{out}$ in V2V attention module is

$$\mathbf{A}^{out} = \boldsymbol{F}(\mathbf{A}^{intra} + \mathbf{A}^{inter}), \tag{7}$$

where $\boldsymbol{F}$ denotes a set of max pooling, average pooling, and convolution layers. For 3D object detection, we use the smooth L1 loss for bounding box regression and focal loss [71] for classification. The final loss is the combination of detection and LC-aware repair loss $\mathcal{L}_{LC}$ as follows,

$$\mathcal{L}_{total} = \mu\mathcal{L}_{det} + \lambda\mathcal{L}_{LC}, \tag{8}$$

where μ and λ are the balance coefficients within range $[0, 1]$.

IV. Experiment

A. Dataset

Due to the difficulties of collecting the real-world CAV perception data for cooperative perception with lossy communication in realistic scenes, we use the digital-twin-based simulation dataset to validate the proposed method. The experiments are conducted on the public cooperative perception dataset OPV2V [10]. OPV2V is a large-scale open-source simulated dataset for V2V perception, which contains 73 divergent scenes with various numbers of connected vehicles, 11,464 frames, and 232,913 annotated 3D vehicle bounding boxes. These data are collected from 8 digital towns in CARLA [16], and a digital town of Culver City, Los Angeles with the same road topology. Following the default setting of OPV2V [10], we use 3,382 frames and 1,920 frames from OPV2V as the training set and validation set, respectively, and 2,170 frames in CARLA Towns and 594 frames in Culver City are used as testing set for all methods.

B. Experiments Setup

Evaluation metrics: We evaluate the performance of our proposed framework by the final 3D vehicle detection accuracy. Following [10], [15], we set the evaluation range as $x \in [-140, 140]$ meters, $y \in [-40, 40]$ meters, where all CAVs are included in this spatial range, whose number is in the range of [1,5] in the experiment. and we measure the accuracy with Average Precisions (AP) at Intersection $-$ over$-$Union (IoU) threshold of 0.5 and 0.7.

Experiment details: In this work, we focus on LiDAR-based vehicle detection and assess models under two scenarios: 1) *Ideal Communication*, where all data transmissions are under perfect communication; 2) *Lossy Communication*, where all intermediate features from other CAVs suffer from the lossy communication except the ego vehicle feature. To simulate the lossy communication, the shared intermediate features are randomly selected by a uniform distributed random probability $p \in [0, 1]$, then replaced by a uniform distributed random noise, which is generated by a uniform distribution within the range of original intermediate features. Statistically, the range of original intermediate features is $[0, 29.5]$ in our experiment.

In the training stage, we adopt two schemes to observe the impact of different training data on V2V 3D object detection models. The *Scheme I* uses only ideal communication-based data for training, while the other *Scheme II* uses simulated lossy communication-based data as described above for training. The training parameter settings for both schemes are identical, and the only difference between them is the training data, which considers lossy communication in *Scheme II* All trained models are evaluated on V2V CARLA Towns and Culver City testing

sets under both *Ideal Communication* and *Lossy Communication* scenarios. Specifically, all models use the PointPillar [37] as the backbone with the voxel resolution of 0.4 m for both height and width. We adopt Adam optimizer [72] with an initial learning rate of 10^{-3} and steadily decay it every 10 epochs using a factor of 0.1. The coefficient of detection loss $\mathcal{L}_{det}$ is set to 1.0, and that of LC-aware repair loss $\mathcal{L}_{LC}$ is set to 0.1. We follow the same hyperparameters in V2X-ViT [15], and all models are trained on two RTX 3090 GPUs.

Compared methods: We consider *No fusion* as the baseline, which only uses the ego vehicle's LiDAR point clouds. In addition, we evaluate five state-of-the-art methods in this paper, which use *Intermediate Fusion* as the main fusion strategy: CoBEVT [11], F-Cooper [12], V2VNet [13], OPV2V [10], and V2X-ViT [15](see SectionII-B for detailed descriptions). To demonstrate the significant effect of Lossy Communication, we first train these methods under two scenarios: *Ideal Communication* and *Lossy Communication*. We then test these methods under the same two scenarios to assess their performance. To show the effectiveness of two critical components in our framework, namely LCRN, and V2VAM, we design a simple feature averaging fusion method with a 1×1 convolutional layer called AveFuse. This method averages all intermediate features from ego-vehicle and other vehicles, and then the averaged feature is passed through a 1×1 convolutional layer.

C. Experimental Results

Table I shows the performance comparisons of all models that are trained with *Scheme I*, then tested on two communication types e.g., *Ideal Communication* and *Lossy Communication*, respectively. Under *Ideal communication*, all the cooperative perception methods significantly surpass *NO Fusion* baseline. In V2V CARLA Town testing set, our proposed V2VAM outperforms the other five advanced fusion methods to achieve 92.6%/86.1% for AP@0.5/0.7, which is highlighted as bold text in Table I. In V2V Culver City testing set, CoBEVT [11] gets 87.7%/74.8% for AP@0.5/0.7, while the V2VAM achieves the 88.5%/78.5% for AP@0.5/0.7 as the best performance, which is higher than the second best fusion method CoBEVT [11] with an AP@0.5/0.7 improvement of 1.6%/3.7%. These results indicate cooperative perception methods can improve the perception performance than a single vehicle perception system under *Ideal Communication*, and our proposed fusion method V2VAM can enhance the interaction between ego vehicle and other vehicles efficiently, which achieves the best performance. However, under *Lossy Communication* testing scenario, all intermediate fusion methods have a drastic performance drop on two testing sets, and the accuracy of these methods is even less than *NO Fusion*. In V2V CARLA Town testing set, the cooperative perception performance of F-Cooper [12], V2VNet [13], OPV2V [10], and CoBEVT [11] decrease by 80.8%, 85.0%, 85.6%, and 82.5% in AP@0.5, respectively. Obviously, all intermediate fusion methods without considering the lossy communication are not practical for deployment in the real world.

The result of 3D object detection on two OPV2V testing sets based on the training of *Scheme II* is presented in Table II.

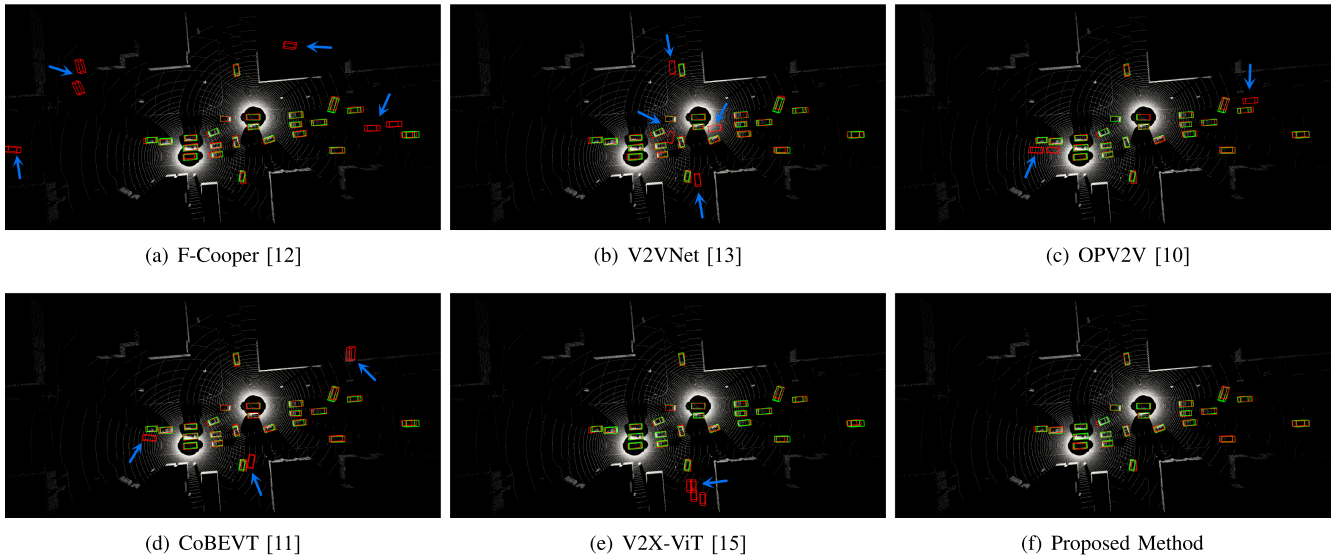
TABLE II
3D DETECTION PERFORMANCE COMPARISON ON TWO TESTING SETS OF OPV2V BASED ON THE TRAINING OF *SCHEME II*

Method	Com. Type	V2V CARLA Towns AP@0.5	V2V CARLA Towns AP@0.7	V2V Culver City AP@0.5	V2V Culver City AP@0.7
NO Fusion	Ideal	0.679	0.602	0.557	0.471
	Lossy	0.679	0.602	0.557	0.471
F-Cooper [12]	Ideal	0.677	0.494	0.758	0.523
	Lossy	0.677	0.492	0.656	0.440
V2VNet [13]	Ideal	0.713	0.465	0.702	0.408
	Lossy	0.714	0.465	0.702	0.409
OPV2V [10]	Ideal	0.804	0.645	0.742	0.576
	Lossy	0.739	0.603	0.718	0.561
CoBEVT [11]	Ideal	0.871	0.740	0.866	0.688
	Lossy	0.768	0.582	0.795	0.586
V2X-ViT [15]	Ideal	0.793	0.619	0.731	0.520
	Lossy	0.770	0.599	0.717	0.511
V2VAM+LCRN	Ideal	**0.887**	**0.783**	**0.871**	**0.709**
	Lossy	**0.841**	**0.705**	**0.846**	**0.663**

Under *Lossy Communication*, although all intermediate fusion methods have a better performance than Table I, which learned the lossy intermediate feature in the training stage. They still fail to handle lossy communication data resulting in the poor perception performance in II. In V2V CARLA Town testing set, F-Cooper [12] got 49.2%, V2VNet [13] got 46.5%, CoBEVT [11] got 58.2%, and V2X-ViT [15] got 59.9% in AP@0.7. These four fusion methods are even worse than single-vehicle baseline *NO Fusion*, which indicates the highly negative impacts by lossy communication. While our proposed method can reach the best performance of 84.1%/70.5% for AP@0.5/0.7 on V2V CARLA town testing set, and 84.6%/66.3% for AP@0.5/0.7 on Culver City testing sets, respectively. The proposed method achieves the best performance under both *Ideal Communication* and *Lossy Communication*, which is highlighted in Table I. Obviously, our proposed LCRN module efficiently maintains the benefits of collaborations under lossy communication. The proposed method can also diminish the impact of lossy V2V communication to achieve excellent cooperative perception performance. Further, we visualize some 3D object detection results on V2V Culver City testing set under *Lossy Communication*, as shown in Fig. 5. Intuitively, these five comparison methods cannot handle loss communication appropriately, thus leading to some false negative proposals. While the proposed method improves the perception performance under lossy communication significantly.

D. Discussion: Different Lossy Communication Types in V2V

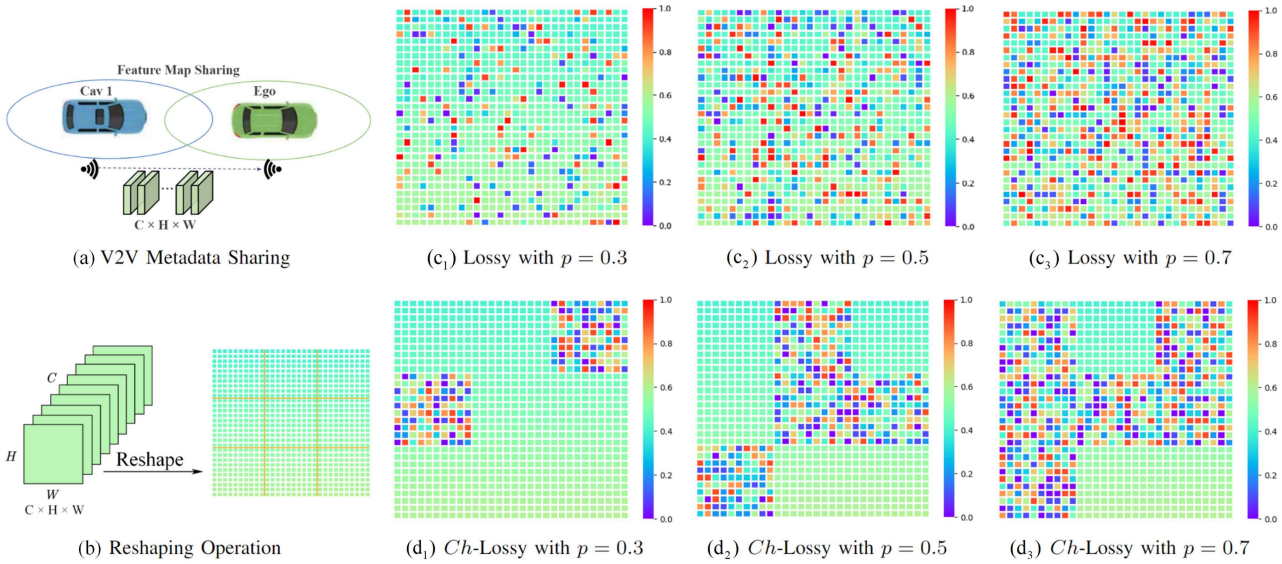
As explained in [56], [73], several random issues such as the occurrence of obstacles, fast and changing vehicle speeds, distance between vehicles might result in lossy communication when sharing a set of communication data. To simulate the complex lossy communication in the real world, the sharing data is randomly selected by a uniformly distributed random probability $p \in [0, 1]$ and then replaced by random noise within the range of original shared feature values. We design two ways



(a) F-Cooper [12] (b) V2VNet [13] (c) OPV2V [10]

(d) CoBEVT [11] (e) V2X-ViT [15] (f) Proposed Method

Fig. 5. 3D object detection visualization. **Green** and **red** 3D bounding boxes represent the **ground truth** and **prediction** respectively. The detection results of the proposed method are clearly more accurate. Some false detection examples are highlighted using blue arrow.



(a) V2V Metadata Sharing (c_1) Lossy with $p = 0.3$ (c_2) Lossy with $p = 0.5$ (c_3) Lossy with $p = 0.7$

(b) Reshaping Operation (d_1) Ch-Lossy with $p = 0.3$ (d_2) Ch-Lossy with $p = 0.5$ (d_3) Ch-Lossy with $p = 0.7$

Fig. 6. Illustration of different lossy communication types in V2V communication. (a) V2V Metadata Sharing, (b) Reshaping Operation, (c_1-c_3) Lossy Communication (named as "Lossy") on the reshaped feature (b) with a global random selection probability p of 0.3, 0.5, 0.7 respectively, (d_1)–(d_3) Channelwise Lossy Communication (named as "Ch-Lossy") on the feature (b) with a channelwise random selection probability p of 0.3, 0.5, 0.7 respectively. We use $C = 9, H = 10, W = 10$ and normalized feature values for illustration in this example.

of random selection to simulate different lossy communication types in the real-world V2V communication.

Lossy Communication (named as "Lossy") on global feature: The shared feature after V2V metadata sharing is reshaped from 3D tensor to 2D matrix first (Fig. 6(b)). Then, as shown in Fig. 6(c_1-c_3), the reshaped feature is randomly selected by the global random probability p and replaced by random noise within the range of original shared feature values.

Channelwise Lossy Communication (named as "Ch-Lossy"): Different with the "Lossy" type to simulate lossy communication on the reshaped global feature, "Ch-Lossy" type is to simulate lossy communication on different channels. As shown in as Fig. 6(d_1)–(d_3), given a shared feature $C \times H \times W$, $\lfloor p * C \rfloor$ channels are randomly selected by the channelwise random probability p and replaced by random noise within the range of original shared feature values.

Finally, the simulated lossy feature is reshaped back to its original shape of $C \times H \times W$ and then received by ego vehicle. In our experiment, *Scheme II* utilizes the simulated lossy communication data by the "Lossy" type to train models, and then we use the models trained in *Scheme II* to test both "Lossy" and "Ch-Lossy" simulated data. Table IV shows the performance

TABLE III
ABLATION STUDY FOR 3D OBJECT DETECTION ON TWO TESTING SETS OF OPV2V BASED ON TRAINING OF *SCHEME II*. NOTE THAT V2VAM+LCRN IS OUR PROPOSED METHOD

Method	V2V CARLA Towns		V2V Culver City	
	AP@0.5	AP@0.7	AP@0.5	AP@0.7
NO Fusion	0.679	0.602	0.557	0.471
AveFuse (Baseline)	0.632	0.325	0.697	0.374
V2VAM *w/o* Intra	0.613	0.490	0.637	0.458
V2VAM *w/o* Inter	0.641	0.494	0.681	0.504
V2VAM	0.709	0.583	0.761	0.541
AveFuse+LCRN	0.698	0.472	0.714	0.558
V2VAM+LCRN	**0.841**	**0.705**	**0.846**	**0.663**

TABLE IV
3D DETECTION PERFORMANCE COMPARISON ON TWO TESTING SETS OF OPV2V BASED ON THE TRAINING OF *SCHEME II* WITH TWO DIFFERENT TYPES OF LOSSY COMMUNICATION

Method	Com. Type	V2V CARLA Towns		V2V Culver City	
		AP@ 0.5	AP@ 0.7	AP@ 0.5	AP@ 0.7
OPV2V [10]	Lossy	0.739	0.603	0.718	0.561
	Ch-Lossy	0.711	0.582	0.737	0.582
CoBEVT [11]	Lossy	0.768	0.582	0.795	0.586
	Ch-Lossy	0.742	0.615	0.767	0.588
V2X-ViT [15]	Lossy	0.770	0.599	0.717	0.511
	Ch-Lossy	0.793	0.619	0.731	0.520
Proposed	Lossy	**0.841**	**0.705**	**0.846**	**0.663**
	Ch-Lossy	**0.852**	**0.723**	**0.851**	**0.675**

comparisons of several methods with the two lossy communication types. The proposed method achieves the best performance under both "Lossy" and "Ch-Lossy" communication types.

E. Ablation Study

The effectiveness of the two proposed components, V2VAM and LCRN, is investigated here. Based on training *Scheme II*, all methods are evaluated under *Lossy Communication* on V2V CARLA Town and Culver City testing sets, respectively. AveFuse is used as the baseline fusion method, which just averages all intermediate features. As shown in Table III, the proposed V2VAM obtains 70.9% in AP@0.5 and 58.3% in AP@0.7 on V2V CARLA Town testing set, which is 7.7% and 25.8% higher than AveFusion in AP@0.5 and AP@0.7 respectively. Both Intra-vehicle attention and Inter-vehicle attention modules are quite effective for V2VAM if we remove one of them in V2VAM during the ablation study. By adding LCRN to the baseline method, AveFuse+LCRN achieves 69.8% in AP@0.5 and 47.2% in AP@0.7 on V2V CARLA Town testing set, with the improvement of 6.6% in AP@0.5, and 14.7% in AP@0.7. Furthermore, our proposed method V2VAM+LCRN achieves the best performance on both V2V CARLA Town and Culver City testing set. Obviously, both V2VAM and LCRN components are beneficial for improving the final performance of 3D object detection in lossy communication scenarios.

V. CONCLUSION

In this paper, the side effect of lossy communication in the V2V cooperative perception is studied, and then we propose the first intermediate LC-aware feature fusion method considering lossy communication. An LC-aware Repair Network (LCRN) is proposed to relieve the side effect of lossy communication and a specially designed V2V Attention Module (V2VAM) is designed to enhance the interaction between the ego vehicle and other vehicles including intra-vehicle attention of ego vehicle and uncertainty-aware inter-vehicle attention. The proposed method is verified in the digital-twin CARLA simulator based public cooperative perception dataset OPV2V, which is quite effective for the cooperative point cloud based 3D object detection under lossy V2V communication and outperforms other V2V point-cloud-based 3D object detection methods significantly.

REFERENCES

[1] R. Xu et al., "The opencda open-source ecosystem for cooperative driving automation research," *IEEE Trans. Intell. Veh.*, early access, 2023, doi: 10.1109/TIV.2023.3244948.

[2] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, "A survey of motion planning and control techniques for self-driving urban vehicles," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 33–55, Mar. 2016.

[3] Y. Ma, C. Sun, J. Chen, D. Cao, and L. Xiong, "Verification and validation methods for decision-making and planning of automated vehicles: A review," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 480–498, Sep. 2022.

[4] D. Cao et al., "Future directions of intelligent vehicles: Potentials, possibilities, and perspectives," *IEEE Trans. Intell. Veh.*, vol. 7, no. 1, pp. 7–10, Mar. 2022.

[5] Q. Lan and Q. Tian, "Instance, scale, and teacher adaptive knowledge distillation for visual detection in autonomous driving," *IEEE Trans. Intell. Veh.*, early access, 2022, doi: 10.1109/TIV.2022.3217261.

[6] K. Strandberg, N. Nowdehi, and T. Olovsson, "A systematic literature review on automotive digital forensics: Challenges, technical solutions and data collection," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1350–1367, Feb. 2023, doi: 10.1109/TIV.2022.3188340.

[7] L. Chen et al., "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1046–1056, Feb. 2023, doi: 10.1109/TIV.2022.3223131.

[8] K. Wang, L. Pu, J. Zhang, and J. Lu, "Gated adversarial network based environmental enhancement method for driving safety under adverse weather conditions," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1934–1943, Feb. 2023, doi: 10.1109/TIV.2022.3142128.

[9] Z. Ju, H. Zhang, X. Li, X. Chen, J. Han, and M. Yang, "A survey on attack detection and resilience for connected and automated vehicles: From vehicle dynamics and control perspective," *IEEE Trans. Intell. Veh.*, vol. 7, no. 4, pp. 815–837, Dec. 2022.

[10] 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. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 2583–2589.

[11] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative Bird's eye view semantic segmentation with sparse transformers," in *Proc. 6th Annu. Conf. Robot Learn.*, 2022. [Online]. Available: https://openreview.net/forum?id=PAFEQQtDf8s

[12] 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. IEEE/ACM Symp. Edge Comput.*, 2019, pp. 88–100.

[13] 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.*, 2020, pp. 605–621.

[14] Y. Tian, J. Wang, Y. Wang, C. Zhao, F. Yao, and X. Wang, "Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 456–465, Sep. 2022.

[15] 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 *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 107–124.

[16] 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.
[17] A. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.
[18] Y. Zhao and S.-G. Haggman, "Intercarrier interference self-cancellation scheme for ofdm mobile communication systems," *IEEE Trans. Commun.*, vol. 49, no. 7, pp. 1185–1191, Jul. 2001.
[19] R. Tavakoli, M. Nabi, T. Basten, and K. Goossens, "An experimental study of cross-technology interference in in-vehicle wireless sensor networks," in *Proc. ACM Int. Conf. Model., Anal. Simul. Wireless Mobile Syst.*, 2016, pp. 195–204.
[20] D. Patra, S. Chavhan, D. Gupta, A. Khanna, and J. J. P. C. Rodrigues, "V2x communication based dynamic topology control in vanets," in *Proc. Int. Conf. Distrib. Comput. Netw.*, 2021, pp. 62–68.
[21] Z. Wang, K. Han, and P. Tiwari, "Digital twin-assisted cooperative driving at non-signalized intersections," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 198–209, Jun. 2022.
[22] Z. Wang et al., "Driver behavior modeling using game engine and real vehicle: A learning-based approach," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 738–749, Dec. 2020.
[23] Z. Hu, S. Lou, Y. Xing, X. Wang, D. Cao, and C. Lv, "Review and perspectives on driver digital twin and its enabling technologies for intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 417–440, Sep. 2022.
[24] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave fmcw radars for perception, recognition and localization in automotive applications: A survey," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 533–555, Sep. 2022.
[25] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain adaptive object detection for autonomous driving under foggy weather," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 612–622.
[26] J. Li, Z. Xu, L. Fu, X. Zhou, and H. Yu, "Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework," *Transp. Res. Part C: Emerg. Technol.*, vol. 124, 2021, Art. no. 102946, .
[27] K. Wang, T. Zhou, X. Li, and F. Ren, "Performance and challenges of 3D object detection methods in complex scenes for autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1699–1716, Feb. 2023, doi: 10.1109/TIV.2022.3213796.
[28] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8555–8564.
[29] D. Rukhovich, A. Vorontsova, and A. Konushin, "Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3 d object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2397–2406.
[30] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2022, pp. 180–191.
[31] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, 2018, Art. no. 3337.
[32] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4490–4499.
[33] L. Fan et al., "Embracing single stride 3D object detector with sparse transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8458–8468.
[34] Y. Wang et al., "Pillar-based object detection for autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 18–34.
[35] T. Y. Chen, C. C. Hsiao, and C.-C. Huang, "Density-imbalance-eased LiDAR point cloud upsampling via feature consistency learning," *IEEE Trans. Intell. Veh.*, early access, 2020, doi: 10.1109/TIV.2022.3162672.
[36] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3D object proposal generation and detection from point cloud," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 770–779.
[37] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12697–12705.
[38] S. Shi et al., "Pv-rcnn: Point-voxel feature set abstraction for 3 d object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10529–10538.
[39] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3D object detector for point cloud," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1951–1960.
[40] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3D object detection from RGB-D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 918–927.
[41] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4604–4612.
[42] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, "Bridging the view disparity between radar and camera features for multi-modal fusion 3D object detection," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1523–1535, Feb. 2023, doi: 10.1109/TIV.2023.3240287.
[43] Y. Li et al., "Deepfusion: Lidar-camera deep fusion for multi-modal 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17182–17191.
[44] J. Nie, J. Yan, H. Yin, L. Ren, and Q. Meng, "A multimodality fusion deep neural network and safety test strategy for intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 6, no. 2, pp. 310–322, Jun. 2021.
[45] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 514–524.
[46] H. Qiu, P.-H. Huang, N. Asavisanu, X. Liu, K. Psounis, and R. Govindan, "Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving," in *Proc. Int. Conf. Mobile Syst., Appl. Serv.*, 2022, pp. 128–141.
[47] Z. Y. Rawashdeh and Z. Wang, "Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2018, pp. 3961–3966.
[48] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.
[49] H. Yu et al., "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21361–21 370.
[50] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "Coopernaut: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17252–17262.
[51] J. Tu, T. Wang, J. Wang, S. Manivasagam, M. Ren, and R. Urtasun, "Adversarial attacks on multi-agent communication," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 7768–7777.
[52] G. Luo, H. Zhang, Q. Yuan, and J. Li, "Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception," in *Proc. Int. Conf. Multimedia*, 2022, pp. 3578–3586.
[53] Z. Lei, S. Ren, Y. Hu, W. Zhang, and S. Chen, "Latency-aware collaborative perception," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 316–332.
[54] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Adv. Neural Inf. Process. Syst.*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, 2022. [Online]. Available: https://openreview.net/forum?id=dLL4KXzKUpS
[55] S. Zeadally, J. Guerrero, and J. Contreras, "A tutorial survey on vehicle-to-vehicle communications," *Telecommun. Syst.*, vol. 73, no. 3, pp. 469–489, 2020.
[56] M. M. Nasralla, C. T. Hewage, and M. G. Martini, "Subjective and objective evaluation and packet loss modeling for 3 d video transmission over lte networks," in *Proc. IEEE Int. Conf. Telecommun. Multimedia*, 2014, pp. 254–259.
[57] P. Watta, X. Zhang, and Y. L. Murphey, "Vehicle position and context detection using v2v communication," *IEEE Trans. Intell. Veh.*, vol. 6, no. 4, pp. 634–648, Dec. 2021.
[58] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X. Wang, "A latency and reliability guaranteed resource allocation scheme for lte v2v communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850–3860, Jun. 2018.
[59] F. Abbas, P. Fan, and Z. Khan, "A novel low-latency v2v resource allocation scheme based on cellular v2x communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 6, pp. 2185–2197, Jun. 2018.
[60] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency v2v communications," in *Proc. IEEE Glob. Commun. Conf.*, 2018, pp. 1–7.

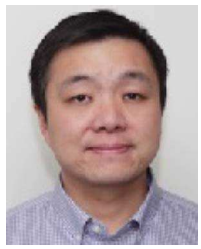
[61] H. Hasrouny, A. E. Samhat, C. Bassil, and A. Laouiti, "Vanet security challenges and solutions: A survey," *Veh. Commun.*, vol. 7, pp. 7–20, 2017.
[62] Y. Yan, H. Du, Q.-L. Han, and W. Li, "Discrete multi-objective switching topology sliding mode control of connected autonomous vehicles with packet loss," *IEEE Trans. Intell. Veh.*, early access, doi: 10.1109/TIV.2022.3215139.
[63] X. Sun, F. R. Yu, and P. Zhang, "A survey on cyber-security of connected and autonomous vehicles (CAVs)," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6240–6259, Jul. 2022.
[64] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3D object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7652–7660.
[65] B. Mildenhall, J. T. Barron, J. Chen, D. Sharlet, R. Ng, and R. Carroll, "Burst denoising with kernel prediction networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2502–2510.
[66] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1712–1722.
[67] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 261–270.
[68] Z. Xia, F. Perazzi, M. Gharbi, K. Sunkavalli, and A. Chakrabarti, "Basis prediction networks for effective burst denoising with large kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11844–11853.
[69] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int.Conf. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6000–6010.
[70] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 603–612.
[71] 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.*, 2017, pp. 2980–2988.
[72] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*., 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7
[73] E. Belyaev, A. Vinel, A. Surak, M. Gabbouj, M. Jonsson, and K. Egiazarian, "Robust vehicle-to-infrastructure video transmission for road surveillance applications," *IEEE Trans. Veh. Technol.*, vol. 64, no. 7, pp. 2991–3003, Jul. 2015.

Jinlong Li (Graduate Student Member, IEEE) received the B.S. and M.S. degrees from Chang'an University, Xi'an, China, in 2018 and 2021, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, deep learning, and intelligent transportation system.

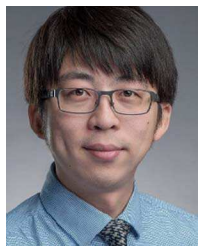
Runsheng Xu received the B.S. and M.S. degrees in electrical engineering from North China Electrical Power University, Beijing, China, and Northwestern University, Evanston, USA, in 2016 and 2017, respectively. He is currently working toward the Ph.D. degree with the UCLA Mobility Lab, University of California, Los Angeles, CA, USA. His research interests include computer vision, autonomous driving perception system, and cooperative driving automation.

Xinyu Liu (Graduate Student Member, IEEE) received the B.S. degree from Polytechnic Institute, Taiyuan University of Technology, Taiyuan, China, in 2017. She is currently working toward the M.S. degree with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH, USA. Her research interests include computer vision, deep learning, and intelligent transportation system.

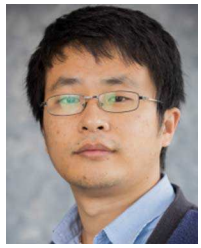
Jin Ma received the B.S. degree in information and computational science and M.E. degree in software engineering from Xi'an Jiaotong University, China, in 2016 and 2019, respectively. He is currently working toward the Ph.D. degree in computer science with Cleveland State University, Cleveland, OH, USA. His research interests include computer vision and deep learning related topics.



Zicheng Chi (Member, IEEE) received the Ph.D. degree in computer engineering from the University of Maryland, Baltimore County, MD, USA. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH, USA. His research interests include communication, fundamental networking, sensing, and energy-related problems for real-world applications in the areas of Internet of Things, and cyber-physical systems.



Jiaqi Ma (Member, IEEE) received the Ph.D. degree in transportation engineering from the University of Virginia, Charlottesville, VA, Virginia, USA, in 2014. He is currently an Associate Professor with the UCLA Samueli School of Engineering and Faculty lead of the New Mobility Program with the UCLA Institute of Transportation Studies. His research interests include intelligent transportation system, autonomous driving, and cooperative driving automation. He is also the Member of the TRB Standing Committee on Vehicle-Highway Automation, TRB Standing Committee on Artificial Intelligence and Advanced Computing Applications, and American Society of Civil Engineers (ASCE) Connected and Autonomous Vehicles Impacts Committee, and the Co-Chair of the IEEE ITS Society Technical Committee on Smart Mobility and Transportation 5.0. He is an Editor in Chief of the IEEE OPEN JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS.



Hongkai Yu (Member, IEEE) received the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Cleveland State University, Cleveland, OH, USA. His research interests include computer vision, machine learning, deep learning, and smart city. He is the Area Chair of ACM Multimedia 2022 and IEEE MIPR 2022 conferences. He is an Editorial Board member of the journal *Green Energy and Intelligent Transportation*.