IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 January 2024, accepted 18 February 2024, date of publication 20 February 2024, date of current version 4 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368404

RESEARCH ARTICLE

ViT-FuseNet: Multimodal Fusion of Vision Transformer for Vehicle-Infrastructure Cooperative Perception

YANG ZHOU[1], CAI YANG[1], PING WANG[1], (Member, IEEE), CHAO WANG[1], (Member, IEEE), XINHONG WANG[1], AND NGUYEN NGOC VAN[2]

[1]College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China
[2]School of Electrical and Electronic Engineering, Hanoi University of Science and Technology, Hanoi 10000, Vietnam

Corresponding author: Ping Wang (pwang@tongji.edu.cn)

This work was supported in part by the Science and Technology Commission of Shanghai Municipality under Grant 22dz1203400 and Grant 21511102400, in part by the International Strategic Innovative Project of National Key Research and Development Program of China under Grant 2023YFE0112500, and in part by the Shanghai Automotive Industry S&T Development Foundation under Grant 2107.

ABSTRACT Perception plays a vital role in autonomous driving as it serves as a prerequisite for downstream planning and decision tasks. Existing research has mainly focused on developing vehicle-side perception models using a single type of sensors. However, relying solely on one type of on-board sensors to perceive the surrounding environment leads to perceptual deficiencies owing to inherent characteristics and sensor sparsity. To address this bottleneck, we propose ViT-FuseNet, a novel vehicle-infrastructure cooperative perception framework that utilizes a Vision Transformer to fuse feature maps extracted from LiDAR and camera data. The key component is a multimodal fusion module designed based on a cross-attention mechanism. ViT-FuseNet has two distinct advantages: i) it incorporates roadside LiDAR point clouds as additional inputs to enhance the 3D object detection capability of the vehicle; and ii) for the effective fusion of data from two different modal sensors, we employ a cross-attention mechanism for feature fusion, rather than directly merging camera features with point clouds at the raw data level. Extensive experiments are conducted using the DAIR-V2X Dataset to demonstrate the effectiveness of the proposed method. Compared with advanced cooperative perception methods, our method achieves a 6.17% improvement in 3D-mAP (IoU=0.5) and an 8.72% improvement in 3D-mAP (IoU=0.7). Moreover, the framework achieves the highest 3D-mAP (IoU=0.5) in all three object categories of benchmarks for single-vehicle perception.

INDEX TERMS Vehicle-infrastructure cooperative perception, multimodal fusion, object detection, vision transformer, cross-attention.

I. INTRODUCTION

Accurate perception of the surrounding environment plays a crucial role in ensuring the safety, efficiency, and robustness of autonomous driving. Downstream tasks, such as planning, decision making and control, rely heavily on upstream perception capability [1]. With the latest advancements in deep learning, the performance of perception algorithms in autonomous driving, such as object detection and tracking, semantic segmentation, and depth estimation, has significantly improved compared with traditional methods applied in the past few years [2]. However, the majority of existing investigations on perception techniques have primarily concentrated on using sensors equipped on a single vehicle. Despite continuous improvements, these methods may not be able to provide sufficiently good performance that satisfies the stringent requirements of autonomous driving applications, owing to factors such as limited sensing range and long-distance occlusion. Increasing the quality and number of onboard sensing and computation equipment may

The associate editor coordinating the review of this manuscript and approving it for publication was Amjad Mehmood.

serve as a solution. But this leads to an additional economical burden and cannot fully solve the problem. Therefore, single-vehicle intelligent perception methods continue to encounter formidable challenges. Therefore, new approaches are desired.

Recently, owing to the rapid development of vehicle-to-everything (V2X) technologies [3], the concept of vehicle-infrastructure cooperative perception, which allows additional information from sensor devices equipped on road-side infrastructure to strengthen single-vehicle perception capability, has attracted widespread research attention in both academia and industry [4]. Infrastructure sensors, such as cameras and LiDARs, are typically mounted at higher positions than those on vehicles, and thus have wider fields of view. They can provide intelligent vehicles with additional sensing resources to enable significant improvements in sensing range and accuracy. Vehicle-infrastructure cooperation holds the potential to revolutionize the entire autonomous driving industry. However, thus far, the research is still in its early stages. For example, unlike the numerous publicly available datasets for single-vehicle perception, there is a lack of high-quality real-world datasets, which are key to training and validating deep learning-based perception algorithms. Most datasets used for vehicle-infrastructure cooperative perception are generated using simulation tools. Recently, the DAIR-V2X dataset [5] collected from practical traffic scenarios was published to support investigations of vehicle-infrastructure cooperative 3D object detection (VIC3D).

More importantly, existing works mainly focus on single-modal perception, that is, conducting sensing tasks based on either camera images or LiDAR point clouds. LiDAR can generate high-resolution point cloud data with high ranging accuracy, which means that it can provide accurate and detailed target position, shape and contour information. This method is useful for object detection, obstacle avoidance, and navigation. However, it lacks rich color and texture information, and is easily affected by atmospheric interference. Conversely, images possess fuzzy depth measurements but offer detailed texture and color information. Exploiting a single type of sensor equipped on the vehicle and infrastructure is relatively straightforward, but it cannot take full advantage of the available multimodal sensing data. Nevertheless, multimodal sensing data fusion is challenging, because the physical world is represented in different ways that are difficult to integrate.

To address this issue, we propose a new framework called ViT-FuseNet, for vehicle-infrastructure cooperative perception using multimodal sensing data. The framework takes the vehicle-side point cloud and image, as well as the roadside point cloud, as inputs to carry out the 3D object detection task. To this end, the point cloud captured by the vehicle LiDAR is first overlaid with that captured by the roadside LiDAR to improve the perception range and quality of the point cloud. Subsequently, the backbone extracts features from the integrated point cloud and vehicle images. Then, the multimodal fusion module fuses the feature maps of the two different modes to generate multimodal feature maps, which can provide rich semantic and spatial information in one feature map. These feature maps are used to obtain the final detection output. The key component of ViT-FuseNet, that is, the multimodal fusion module, is established based on the cross-attention mechanism, which utilizes the vision Transformer (ViT) to unify the features of two different modalities into the same representation space, enabling effective cross-modal feature fusion. Moreover, compared with traditional convolutional networks, the attention mechanism can find the relationship between different modalities based on global features in the early stages, thereby enhancing the representation effectiveness of fusion features [6].

Extensive experiments were conducted on the DAIR-V2X dataset [5] to evaluate the performance of ViT-FuseNet by comparing it with the benchmarks of the dataset and several advanced cooperative perception methods, including V2VNet [7] and DiscoNet [8]. The results show improvements of 17.16% and 23.71% in 3D-mAP (Intersection-over-Union (IoU)=0.5) over the early- and late-fusion LiDAR detection benchmarks, respectively. With sufficiently good V2X transmission and well-synchronized vehicle-side and infrastructure-side point clouds, ViT-FuseNet achieves an improvement of 6.17% in 3D-mAP (IoU=0.5) compared with the advanced cooperative perception methods. The proposed framework also outperforms the baselines for single-vehicle detection. Additional experiments were conducted on the multimodal fusion module to reveal the critical role of position embedding in learning global information when using a ViT for multimodal fusion. Multimodal fusion with cross-attention is also found to capture more spatial information and exhibit better performance in detecting small objects. Thus, the advantages of the proposed framework are clearly demonstrated.

Our main contributions are summarized as follows:

- We propose a vehicle-infrastructure cooperative perception framework for the VIC3D problem, called ViT-FuseNet. It accommodates inputs from heterogeneous sensor devices and facilitates end-to-end detection in vehicle-infrastructure cooperative scenarios to enhance the accuracy of 3D object detection.
- We introduce a novel vision Transformer module that employs a cross-attention mechanism for multimodal feature fusion. This module can effectively combine data from different modal sensors while capturing the interactions of features in both adjacent and global spatial domains. This ensures that the fused features provide rich and valuable information for cooperative perception tasks.
- We evaluate the proposed ViT-FuseNet framework on the real-world DAIR-V2X dataset. Superior performance was achieved compared with several state-of-the-art object detection methods.

II. RELATED WORK

Three-dimensional (3D) object detection is an important branch of computer vision. It utilizes sensors, such as cameras and LiDAR installed on vehicles, along with corresponding perception algorithms, to detect traffic participants. However, with the high cost of single-vehicle perception, attention has gradually shifted towards vehicle-infrastructure cooperative perception technology. It uses both roadside infrastructure and on-board perception devices to perform object detection tasks, expanding the field of view while sharing the cost burden. It also introduces new challenges, such as the fusion of heterogeneous perception devices. The ViT technology provides a new approach for multimodal perception. It employs the attention mechanism to capture the correlation information between the adjacent and global features.

A. 3D OBJECT DETECTION IN AUTONOMOUS DRIVING

Object detection plays a crucial role in autonomous driving perception. Its accuracy limits the quality of downstream tasks such as planning and decision-making. Based on the sensor types, 3D object detection can be classified into three categories: camera-based methods, LiDAR-based methods, and multi-sensor-based methods. An example of the camera-based method is ImVoxelNet [9], which is a novel convolutional method based on posed monocular or multi-view RGB images. Another example is the BEVformer [10], which projects 2D images onto the bird's-eye view (BEV) to perform multi-camera-based 3D detection. LiDAR-based methods are the most widely used approach for autonomous driving. PointNet [11] and PointNet++ [12] are pioneering studies that directly apply neural networks to point clouds. Two other 3D detection methods emerged later, i.e., the voxel-based and pillar-based methods. VoxelNet [13] discretizes the point cloud into a 3D grid, and then a 3D convolutional detection network is applied. SECOND [14] investigated an improved sparse convolution and a new form of the angle-loss regression method, which significantly increases the speed of both training and inference. Because processing 3D voxels is often computationally expensive, pillar-based methods, such as PointPillars [15], have been developed to convert voxels with the same z-axis into a 2D pillar representation in the BEV for faster feature processing. In this study, we select PointPillars as the backbone for processing LiDAR point clouds because of its fast inference speed and low memory usage.

Multi-sensor-based methods, such as Pointpainting [16] and MVXNet [17], utilize both camera and LiDAR data. Images typically provide rich semantic and texture information, whereas LiDAR point clouds provide clear distance and depth information. Ideally, integrating the two types of sensor data should yield better results than using only one type. However, early works on leveraging the synergy of cameras and LiDAR, such as MV3D [18], may lead to even worse performance than the algorithms that only utilize LiDAR point clouds (e.g., PointPillars [15] and STD [19]) to perform detection. This is because it is difficult to effectively align and integrate data with different representations (such as LiDAR point cloud and camera image data) to obtain a unified semantic feature space, which is also the focus of this study.

B. COOPERATIVE PERCEPTION

As an import application of the V2X communication technology, multi-agent cooperative perception has started to attract research attentions. For instance, V2VNet [7] applies a graph neural network-based multilayer information iteration to integrate sensing data collected from multiple vehicles to achieve better perception performance. When2com [20] saves channel resources through handshake communication mechanisms to ensure real-time perception. DiscoNet [8] adopts knowledge distillation to leverage the advantages of early fusion (raw data fusion) and middle fusion (feature fusion). OPV2V [21] proposed a graph-based self-attention feature fusion to enhance perception performance, along with V2X-Sim [22], serving as two simulated datasets for research on multi-vehicle cooperative perception. V2X-ViT [23] introduced a heterogeneous multi-agent attention module to fuse sensing information among different types of intelligent agents. These pioneering works are conducted based on customized or simulated data, which may not be able to fully reflect the issues presented in real-world scenarios, such as imperfect temporal and spatial data alignment, limited communication bandwidth, and non-negligible transmission latency. Thus, comprehensive investigation and resolution using real-world data is necessary.

Recently, DAIR-V2X [5] releases a large-scale real-world dataset, collected by both vehicle and roadside LiDAR and cameras, for vehicle-infrastructure cooperative 3D object detection (VIC3D). Several baseline algorithms for early and late fusion are also provided. Based on this dataset, FFNet [24] develops an intermediate feature fusion framework that addresses the challenge of temporal asynchrony between the input of the framework. However, the majority of existing studies on VIC3D still focus on single-modal sensing fusion, mainly using point clouds. The benefits of perception with LiDAR and camera fusion have not been fully exploited. We intend to fill inthis knowledge gap.

C. VISION TRANSFORMER

Transformer [6] was initially developed for machine translation. Its advantages lie in the use of the multihead self-attention mechanism and feed-forward networks to capture long-range interactions between words. Dosovitskiy et al. [25] proposed a Vision Transformer (ViT) to unify the fields of computer vision and natural language processing (NLP), by dividing an image into multiple image patches, each being treated as a visual token. To address the challenge of heavy computational complexity that hinders scalability to long sequences or high-resolution images, several methods introduce locality into self-attention, such as Swin [26] and

CSwin [27]. Typically, they adopt a hierarchical structure to progressively increase the receptive field and capture longer dependencies. The core idea of the attention mechanism is to selectively identify and focus on a small amount of crucial information relevant to the current task, while ignoring less significant information. The ViT framework with a self-attention mechanism has proven to be effective in modeling homogeneous structured data. However, it has not yet been widely applied for the representation modeling of multimodal heterogeneous data in VIC3D tasks.

To address the above issues, in this study, we propose a ViT-based solution for vehicle-infrastructure cooperative multimodal fusion perception that effectively integrates camera and LiDAR data with different representations. Our method can learn to capture the correlation between image features and point cloud features through a cross-attention mechanism and unify the information from two modalities into a common feature map. In addition, it enhances the vehicle-side sparse point cloud through the infrastructure point cloud and improves perception performance. The middle fusion approach can significantly reduce the computation and communication resources. Details of the proposed framework are presented in the following section.

III. METHOD

We consider a scenario in which an intelligent vehicle equipped with LiDAR and a camera drives close to a roadside infrastructure with LiDAR. The three sensors have an overlapping field of view (OFV). We aim to develop a multimodal feature fusion framework to realize VIC3D and enhance the performance of vehicle-side detection. To achieve this goal, PointPillars [15] is employed as the backbone for LiDAR point cloud feature extraction. PointPillars was originally developed for object detection by using only point clouds. However, this may result in limited detection accuracy owing to the lack of semantic and color information in the LiDAR data. To improve the performance, ViT-FuseNet incorporates a Transformer encoder based on the cross-attention mechanism, effectively integrating features from LiDAR point clouds and camera images. The fused feature map contains rich semantic and depth information, which effectively compensates for the sparse LiDAR point cloud features and improves detection performance.

The overall architecture of our framework is illustrated in Fig. 1, and includes four major phases: 1) data sharing, 2) feature extraction, 3) ViT multimodel fusion, and 4) detection head. The framework accommodates inputs from heterogeneous sensor devices and facilitates end-to-end detection in vehicle-infrastructure cooperative scenarios to enhance the accuracy of 3D object detection. The key component of our framework is the third part: the ViT multimodal fusion module. It can effectively combine data from different modal sensors while capturing the interactions of features in both adjacent and global spatial domains. In the following sections, we describe the main architecture design of the framework in Section III-A and the details of the ViT multimodal fusion module in Section III-B.

A. MAIN ARCHITECTURE DESIGN

1) DATA SHARING

ViT-FuseNet aims to integrate the roadside LiDAR and vehicle-side camera sensing data with the vehicle-side LiDAR point cloud to carry out VIC3D. We assume that the V2X communication between the vehicle and infrastructure is of sufficient quality for data exchange to be realized with negligible error and delay. In the first phase of the framework, the infrastructure attains the real-time location of the vehicle, and determines the OFV. It then projects its LiDAR point cloud in the OFV onto the vehicle coordinate system using a coordinate transformation matrix, and transmits it to the vehicle. The vehicle overlays the received roadside infrastructure point cloud with its own LiDAR point cloud, and determines the region of multimodal fusion based on the perception range of the camera.

2) FEATURE EXTRACTION

Upon reception of the infrastructure sensing data, the second phase of ViT-FuseNet performs feature extraction. Two networks are adopted: one for image features and one for point cloud features. The image branch uses a semantic segmentation network for feature extraction to capture abundant colors and semantic features. In our study, we use the output of the last DSConv layer of a pre-trained Fast-SCNN network [28] as the original image feature map. The feature map is upsampled to the original image size and aligned with the dimension of the LiDAR point cloud features, generating the image feature map, denoted by $F_i \in \mathcal{R}^{H\times W\times C}$ with H is the height of the feature map, W is the width, C is the number of channels, and $\mathcal{R}$ is the set of real numbers.

For the point cloud branch, we leverage the anchor-based PointPillars backbone [15] to extract features because of its rapid inference speed and optimized memory usage. The integrated point cloud is first converted to a stacked pillar tensor and then scattered to a 2D pseudo-image and fed into the PointPillars backbone. The backbone generates the informative feature map $F_p \in \mathcal{R}^{H\times W\times C}$. F_i and F_p represent the features of the two different modalities.

3) VIT MULTIMODAL FUSION

The intermediate features F_i and F_p, which are extracted from the camera and LiDAR, are input to the key component of our framework, namely the ViT multimodal fusion module. The Transformer encoder in this module employs a cross-attention mechanism to iteratively learn the correlation between the image and point cloud features, thereby facilitating their interactive fusion. The powerful modeling capability of ViT enables the generation of a multimodal feature map, denoted as $F_m \in \mathcal{R}^{H\times W\times C}$. The

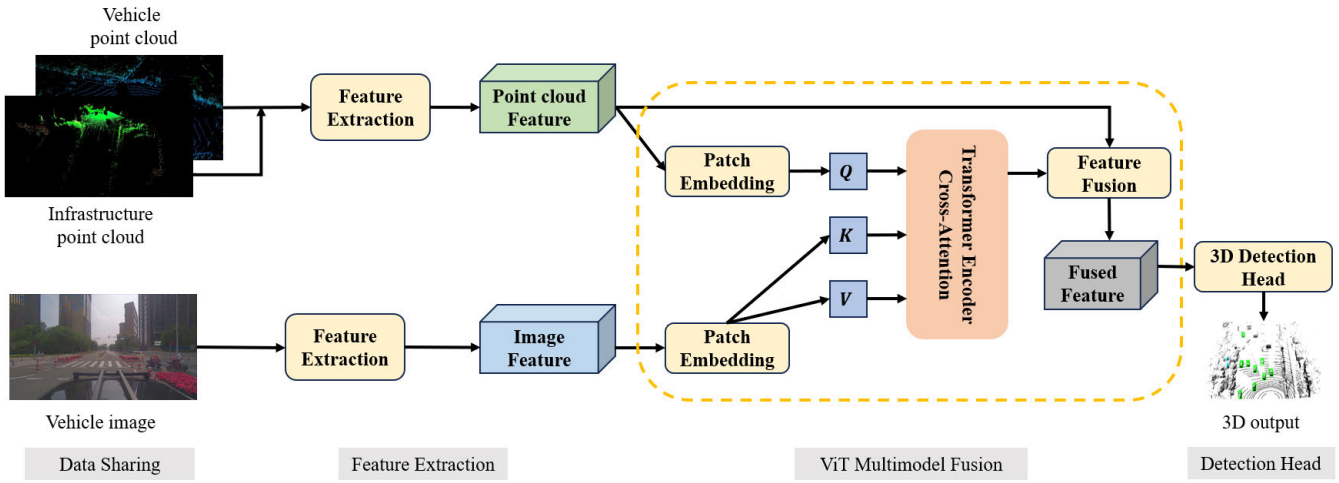


FIGURE 1. The ViT-FuseNet framework and the four major operation phases: 1) Data sharing: Attain LiDAR point cloud from the infrastructure through V2X communication, 2) Feature extraction: Extract image and point cloud features, 3) ViT multimodal fusion: Employ the cross-attention mechanism to learn interactions between multimodal features and fuse them in the same representation space, 4) Detection head: Apply a detection head for object predictions.

implementation details of this key module are explained in Section III-B.

It is important to note that, compared with the traditional method of simply splicing two different modal features, our fusion module has no obvious bias to the two modal information. The model dynamically determines the key features to be fused based on the desired outcomes. Meanwhile, considering that the cross-attention mechanism requires explicit Query, Key, and Value matrices, we use point cloud feature as Query and image feature as Key and Value. This enables the model to query image features that can significantly enhance the point cloud features for positive correlation fusion.

4) DETECTION HEAD

The final phase involves performing the 3D object detection using the output of the ViT multimodal fusion module F_m. Two convolution layers are employed for box regression and classification. We adopt a Single Shot Detector (SSD) [29] as the 3D object detection head to generate 3D outputs for more accurate localization and recognition.

B. VIT MULTIMODAL FUSION

Our goal is to design a customized ViT module that can handle the common challenges of multimodal feature fusion in vehicle-infrastructure cooperative perception tasks. The structure of the module is illustrated in Fig. 2(a). It consists of three sequential steps. First, to effectively utilize the Transformer to fuse point cloud and image features, the patch embedding step is carried out to process the feature map into an appropriate dimension. Subsequently, a novel Transformer encoder formed by a MultiHead Cross-Attention (MCA) block and a multilayer perceptron (MLP) block is employed to generate image-enhanced features. Finally, the output of the Transformer encoder goes into the feature fusion step, which effectively integrates the image-enhanced features with the original point cloud features. These methods significantly improve the effectiveness of feature fusion for different representation forms, leading to a robust aggregated feature representation for detection. The detailed design of each step is as follows.

1) PATCH EMBEDDING

The standard Transformer uses a 1D sequence of token embeddings as the input. The image and point cloud feature maps obtained through their feature extraction networks are both 2D. To produce proper inputs to the Transformer encoder, we implement patch embedding on the image feature map F_i and the point cloud feature map F_p respectively. The procedure for each feature map is illustrated in Fig. 3. We reshape both the image and point cloud feature map into a sequence of flattened 2D patches $x_p^i \in \mathcal{R}^{L\times(P^2\times C)}$, where i is the serial number, C (64) is the number of channels, $P \times P$ (16×16) is the resolution of each feature patch, and $L = HW/P^2$ (837) is the resulting number of patches. Next, we flatten the patches and map them to D (256) dimensions with a trainable linear projection matrix $E \in \mathcal{R}^{(P^2\times C)\times D}$ to accommodate the multihead attention mechanism as

$$z_0 = [x_p^1 E; x_p^2 E; \cdots; x_p^N E] + E_{pos}, \tag{1}$$

where $E_{pos} \in \mathcal{R}^{L\times D}$ is the position embedding introduced in the next paragraph. Finally, both the image and point cloud feature sequences have dimensions of $L \times D$ as z_0, where L serves as the effective input sequence length for the Transformer encoder and each patch within the sequence has a feature vector with a length of D. It is noteworthy that Eq. (1) is a normal form operation for both the image and the point cloud feature maps.

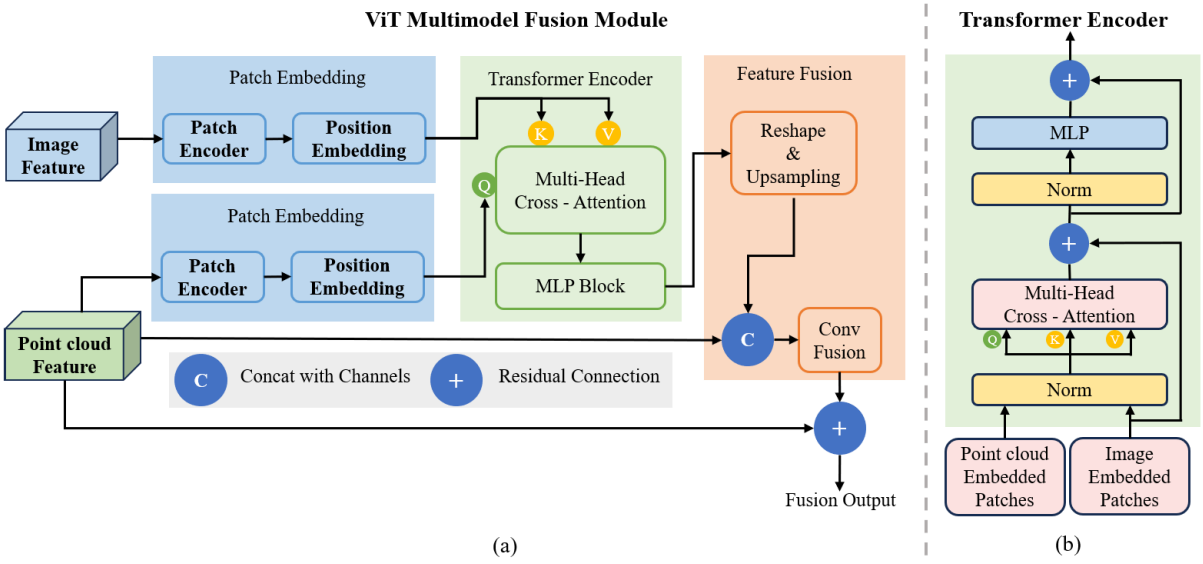


FIGURE 2. (a) The structure of the ViT multimodal fusion module, and (b) the Transformer Encoder. The fusion module consists of three sequential steps: patch embedding, Transformer encoder and feature fusion. The main blocks of Transformer encoder are MultiHead Cross-Attention and MLP blocks.

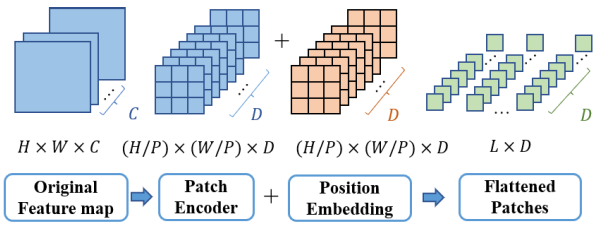


FIGURE 3. The Patch Embedding procedure: We split a feature map into fixed-size patches, linearly embed each of them, add position embedding, and feed the resulting sequence of vectors to the Transformer encoder.

The Transformer encoder simultaneously processes all patches in the sequence, which means that the model itself does not consider the order of each patch. However, context correlation typically exists between adjacent feature patches [25]. Moreover, our research in Section IV-C suggests that positional information between the feature patches is essential. Therefore, position embeddings E_{pos} are added to the patch embeddings to retain positional information, as shown in Fig. 3. We employ standard learnable 1D position embedding [25] because we do not observe notable performance improvements when using more advanced 2D-aware position embedding in Section IV-C, which agrees with the conclusion of Dosovitskiy et al. [25]. The resulting sequence of embedding vectors z_0 as shown in Eq. (1) from LiDAR and camera feature maps serves as the input to the Transformer encoder.

2) TRANSFORMER ENCODER

This step is responsible for capturing the correlation and cross-modal interactions between the encoded point cloud features and image feature sequences. The overall architecture of our encoder, which mainly consists of a MCA block and a MLP block, is illustrated in Fig. 2(b). The output of the former, denoted as z'_f, can be expressed as

$$z'_f = \mathrm{MCA}\left(\mathrm{LN}\left(z_0^i, z_0^p\right)\right) + z_0^i, \tag{2}$$

where z_0^i and z_0^p are the sequences of feature vectors from the images and point clouds, respectively. The output of the MLP block, denoted by z_f, is

$$z_f = \mathrm{MLP}\left(\mathrm{LN}\left(z'_f\right)\right) + z'_f, \tag{3}$$

which is also the output of the Transformer encoder. Moreover, layer normalization (LN) is applied before each block, and residual connections are employed after each block to facilitate the learning of identity mapping by the model and mitigate the issue of gradient vanishing.

Standard cross-attention is a widely-used building block in deep neural network architectures. It can capture the interrelationship between different modalities and long-range interactions at various scales [23], thus, we hope to utilize its ability to extract image features that can enhance point cloud features. For each feature vector in the input sequence z_0 shown in Eq. (1), we employ three trainable linear layers with different weight matrices to obtain query Q, key K, and value V. Specifically, Q is projected from the point cloud feature sequence, whereas K and V are projected from the image feature sequence. The weight matrices of the three linear layers, W_q, W_k, and W_v, are initialized with random Kaiming initialization [30] to increase the robustness. The resulting Q, K, and V serve as inputs for the MCA block shown in Fig. 2(b), with dimensions of $L \times D$, matching the original sequence dimensions.

In the MCA block, we compute the attention weights based on the pairwise similarity between the features in the

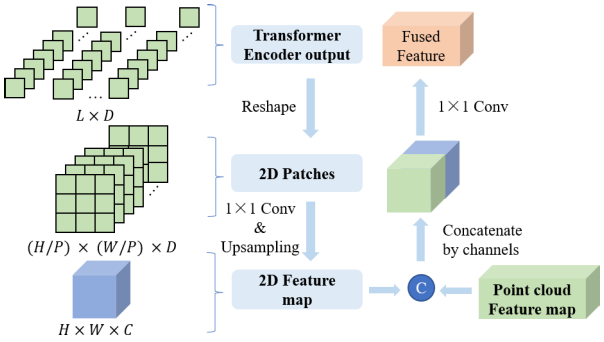


FIGURE 4. The Feature Fusion procedure: The attention features generated by the Transformer encoder go through a series of processing, and then are fused with the original point cloud features.

sequence and their corresponding Q and K representations. Subsequently, we use these attention weights to compute a weighted sum over all V in the sequence as

$$\text{Cross-Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right) \cdot V, \quad (4)$$

where K^T is the transpose of matrix K. This operation allows us to obtain a weighted image feature matrix that can enhance the point cloud features.

MCA is an extension of the cross-attention (CA) where we run k parallel cross-attention operations, called ''heads'', and project their concatenated outputs. It enables the model to learn feature representations from different subspaces, effectively attend to features of various aspects of the input, capture richer contextual information, and enhance model expressiveness and perceptual capabilities [31]. To maintain consistent computation and parameter counts when varying k, the number of feature map channels D of each set is usually set to D_h by D/k. At this stage, the dimensions of Q, K, and V are transformed into $k \times L \times (D/k)$. Each set z_i, including Q_i, K_i, V_i, has dimensions of $L \times D/k$. The CA is applied within each set as

$$\text{MCA}\,(z_0) = \text{Concat}\,(\text{CA}\,(z_1), \text{CA}\,(z_2), \ldots, \text{CA}\,(z_k)), \quad (5)$$

where z_i, $(i = 1, 2, \ldots, k)$ is each set of z_0 in Eq. (1). The results of the k sets are concatenated along the channels to generate the output of the MCA block. Subsequently, the encoder module produces the final output through residual connections, LN and an MLP block as shown in Eq. (3).

3) FEATURE FUSION

The Transformer encoder module generates weighted image sequence features that can enhance the point cloud features. It is necessary to restore these weighted image sequence features to match the size of the original feature map and integrate them with the original point cloud features. As depicted in Fig. 4, the dimensions of the image sequence features are $L \times D$. Initially, the sequence features are stacked to reshape a 2D feature map with dimensions of $(H/P) \times (W/P) \times D$. Subsequently, a 1×1 convolution is employed to restore the dimensions of the feature map to $(H/P) \times (W/P) \times C$, with the same number of channels as the original features. Bilinear upsampling is then utilized to resize the feature map from $(H/P) \times (W/P)$ to $H \times W$, generating weighted image features with dimensions of $H \times W \times C$. Finally, the weighted image features are concatenated with the original point cloud features along the channels, and 1×1 convolution is used for channel fusion to reshape the number of channels to C. The procedure can be expressed as

$$F_m = \text{Conv}\left(\text{Upsample}\left(\text{LN}\left(z_f\right)\right) \oplus F_p\right) + F_p, \quad (6)$$

where z_f is the output of the Transformer encoder and F_p is the original point cloud feature from the beginning. The resulting fused feature F_m with rich image and point cloud information is fed into the detection head to accomplish the 3D object detection tasks.

IV. EXPERIMENTS AND PERFORMANCE EVALUATION

In this section, we use the DAIR-V2X dataset [5] to evaluate the performance of our ViT-FuseNet framework for the VIC3D task. We compared the proposed method with several 3D object detection methods that utilize different fusion strategies. The experimental results demonstrate that our solution outperforms other existing methods, supported by sufficiently good V2X communication. Furthermore, we examined the impact of position embedding in the ViT multimodal fusion module on the perception performance. The detection results are all evaluated by the mean Average Precision (mAP) at an intersection-over-union (IoU) threshold of 0.50 and 0.70, respectively.

A. DATASET AND EXPERIMENTAL SETUP

The DAIR-V2X dataset [5] is a large-scale real-world vehicle-infrastructure cooperative perception dataset composed of images and point clouds captured by cameras and LiDARs mounted on both vehicles and roadside infrastructure. It covers over 100 traffic scenarios collected in the Beijing Advanced Autonomous Driving Demonstration Zone. The dataset is partitioned into three subsets: the vehicle-infrastructure cooperative dataset (DAIR-V2X-C), the infrastructure dataset (DAIR-V2X-I), and the vehicle dataset (DAIR-V2X-V). The DAIR-V2X-C dataset comprises 9,311 pairs of vehicle-infrastructure data frames, offering cooperative annotations (only for one single class ''Car'') from both vehicle and infrastructure perspectives. The DAIR-V2X-V dataset has 22,325 data frames, with labels on three different classes of objects, including ''Car'', ''Pedestrian'', and ''Bicycle''.

Our model is trained and tested using the PyTorch deep learning framework in a Linux environment on an NVIDIA GeForce RTX 3090 GPU. PointPillars [15] is used as the backbone network. Hence, the network parameter setup and training strategy are chosen mainly following [15]. Specifically, the parameters used in the ViT multimodel fusion module presented in Section III-B are listed in Table 1. It is worth noting that these parameters can be modified and

TABLE 1. The network parameters of the ViT multimodal fusion module.

Network module	Layer	Parameters	Input Size	Output Size
Patch Embedding	Conv2d	Kernel size: 16 × 16 Stride: 1; Padding: 1 Kernel number: 256	496×432×64	31×27×256
	Position Embedding	/	31×27×256	31×27×256
	Flatten Layer	/	31×27×256	837×256
Transformer Encoder	Singlehead2Multihead	Head number: 4; dim: 64	837×256	4×837×64
	Multihead Cross-Attention	/	4×837×64	837×256
	MLP	Dropout: 0.4 FC1: 256,128 FC2: 128,256	837×256	837×256
Convolutional Fusion	Flatten Layer	/	837×256	31×27×256
	Conv2d	Kernel size: 1 × 1 Stride: 1; Padding: 0 Kernel number: 64	31×27×256	31×27×64
	Upsample	/	31×27×64	496×432×64
	Concat	/	496×432×64	496×432×128
	Conv2d	Kernel size: 1 × 1 Stride: 1; Padding: 0 Kernel number: 64	496×432×128	496×432×64

should vary based on factors such as the number of attention heads and position embedding dimensions. In our work, to balance performance and memory usage, the dimensions of the original input feature maps F_i and F_p, that is, $H \times W \times C$, are set to be $496 \times 432 \times 64$, and the number of sequence features is set to $D = 256$. The number of attention heads for the MCA block is chosen to be $k = 4$, resulting in each set feature having $D/k = 64$ channels.

The training process for the proposed model consists of 90 epochs. The initial learning rate is set to 0.0003 and the maximum learning rate is set to 0.003. Throughout the training, the learning rate follows a cosine function to control the step size, initially increasing and then decreasing. The ascending phase accounts for 40% of the entire training process to enhances robustness. The detection ranges of the vehicle-side point cloud on the x, y, z axes are constrained to be $(-39.68, 39.68)$, $(0, 69.12)$, $(-3, 1)$ meters, respectively. The voxel sizes for the x, y, and z axes are 0.16, 0.16, and 4 meters, respectively. All input data are pre-converted to the KITTI format [32], following the guidelines specified in [5]. They are subsequently divided into training, validation, and test sets at a $5 : 2 : 3$ ratio. Evaluation is performed on the validation set.

B. PERFORMANCE EVALUATION

We compare our method with several existing models for 3D object detection. They cover a wide range of cooperative perception and sensing fusion strategies, including independent perception (i.e., detection with data from sensors mounted on a single entity), cooperative perception (i.e., detection with data from sensors mounted on multiple entities), no fusion (i.e., detection with a single sensor), early fusion (i.e., raw data fusion), middle fusion (i.e., feature fusion), late fusion (i.e., detection result fusion), and single-modal fusion (that is, fusion of LiDAR data only), and multimodal fusion (i.e., fusion of LiDAR and camera data). To ensure a fair comparison, all models that perform the detection task using the LiDAR point cloud utilize the PointPillars model as the backbone.

1) VIC3D OBJECT DETECTION

We first compare the performance of different models in conducting the 3D object detection task. Independent perception solutions are applied to the vehicle-side sensing data from the DAIR-V2X-C dataset. For cooperative perception approaches, infrastructure-side data are further used to enhance the vehicle-side detection performance. Table 2 presents the results of this evaluation. Only the class ''Car'' is taken into account, because dataset only provides cooperative annotations of the class ''Car''.

These methods include three baseline methods from the DAIR-V2X benchmarks: non-cooperation (e.g., PointPillars [15]), early fusion, and late fusion (e.g., TCLF [5]). Additionally, we compare it with the existing advanced cooperative perception method DiscoNet [8], large-scale vehicle-to-vehicle cooperation network V2VNet [7], feature fusion baseline FFNet [24] which aims to address latency and localization errors, and vehicle-infrastructure cooperative multimodal perception network Multistage Fusion [34]. The detection performance is measured using KITTI [32] evaluation detection metrics: BEV-mAP and 3D-mAP with 0.5 IoU and 0.7 IoU, respectively. Clearly, the proposed ViT-FuseNet framework exhibits better performance than the other solutions.

Specifically, compared with the single-vehicle perception baseline PointPillars [15], our method shows a significant improvement of 31.71% in 3D-mAP (IoU=0.5) and 29.5% in BEV-mAP (IoU=0.5). These results reveal the advantages of conducting vehicle-infrastructure cooperative multimodal perception in future autonomous driving systems.

TABLE 2. VIC3D object detection performance comparison, on the DAIR-V2X-C dataset. The performance of the F-Cooper, V2VNet, DiscoNet, OPV2V is provided in [35]. "-" means that the associated results are not reported. ViT-FuseNet-V means only using vehicle point clouds as LiDAR input without vehicle-infrastructure cooperation.

Model	Cooperation	Fusion Type	Sensors	3D-mAP		BEV-mAP
				IoU=0.5	IoU=0.7	IoU=0.5
Pointpillars [15]	Vehicle-only perception	Non-fusion	LiDAR	48.06	-	52.24
Early Fusion [5]	Cooperative perception	Early	LiDAR	62.61	-	68.91
Late Fusion [5]	Cooperative perception	Late	LiDAR	56.06	-	62.06
F-Cooper [33]	Cooperative perception	Middle	LiDAR	73.40	55.90	-
V2VNet [7]	Cooperative perception	Middle	LiDAR	66.40	40.20	-
DiscoNet [8]	Cooperative perception	Middle	LiDAR	73.60	58.30	-
OPV2V [21]	Cooperative perception	Middle	LiDAR	73.30	55.30	-
FFNet [24]	Cooperative perception	Middle	LiDAR	55.81	30.23	63.54
Multstage Fusion [34]	Cooperative perception	Middle & Late	LiDAR & Camera	71.81	-	78.96
ViT-FuseNet-V(Ours)	Vehicle-only perception	Middle	LiDAR & Camera	78.84	65.20	81.47
ViT-FuseNet(Ours)	Cooperative perception	Early & Middle	LiDAR & Camera	**79.77 (+6.17)**	**67.02 (+8.72)**	**81.74 (+2.78)**

TABLE 3. Performance gained by ViT multimodal fusion module on the DAIR-V2X-V dataset. The performance of the ImVoxelNet, Pointpillars, SECOND, and MVXNet models is provided in [5].

Model	Modality	$Car_{3D-mAP(IoU=0.5)}$			$Pedestrain_{3D-mAP(IoU=0.25)}$			$Cyclist_{3D-mAP(IoU=0.25)}$		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
ImVoxelNet [9]	Image	56.86	39.74	33.00	9.09	9.09	9.09	10.48	9.09	9.09
PointPillars [15]	Point cloud	61.76	49.02	43.45	33.40	24.68	22.39	38.24	33.80	32.35
SECOND [14]	Point cloud	69.44	59.63	57.63	43.45	39.06	38.78	44.21	39.49	37.74
MVXNet [17]	Image & Point cloud	69.86	60.74	59.31	47.73	43.37	42.49	45.68	41.84	40.55
Ours	Image & Point cloud	**72.78 (+2.92)**	**64.52 (+3.78)**	**60.15 (+0.84)**	**59.60 (+11.87)**	**49.85 (+6.48)**	**49.57 (+7.08)**	**50.99 (+5.31)**	**48.99 (+7.15)**	**48.53 (+7.98)**

Additionally, compared with the early fusion and late fusion benchmarks provided by DAIR-V2X [5], the proposed ViT-FuseNet achieves an improvement of 17.16% in 3D-mAP (IoU=0.5) and 12.83% in BEV-mAP (IoU=0.5). Such observations confirm that camera sensing data play a positive role in supporting point cloud data in perception tasks, and the potential of multimodal information should be taken into consideration. The multimodal fusion module proposed in this study, which is based on the cross-attention mechanism, effectively integrates features from two distinct representation forms, resulting in efficient fusion and enhanced perception performance.

F-cooper [33], V2VNet [7], DiscoNet [8], OPV2V [21], and FFNet [24][1] have recently emerged as 3D object detection models via feature fusion. In comparison with these advanced perception methods, our ViT-FuseNet framework achieves varying degrees of improvement, ranging from 6.17% to 23.96% in 3D-mAP (IoU=0.5). The observations show that the effective fusion of data from two different modalities of features through the cross-attention mechanism enables significant enhancement of LiDAR sensors. And the rich color and texture information from image features are extracted to facilitate LiDAR-based perception methods to effectively reduce instances of missed detection and false perception. Furthermore, the point cloud of the infrastructure LiDAR offers abundant information for vehicle perception, compensating for the sparsity of the vehicle point cloud.

In addition, a Multistage Fusion method is developed in [34] to carry out vehicle-infrastructure cooperative multimodal perception. The vehicle-side utilizes a two-stage detection network to fuse the feature maps of images and point clouds, which are then combined with the infrastructure detection results to generate the final detection outcomes. When compared with the baselines of the early fusion and late fusion methods [5], the Multistage Fusion method exhibits significant performance improvements, as shown in Table 2. This demonstrates the superiority of multimodal information in the field of object detection. However, it does not surpass all LiDAR-based feature fusion methods. Our method, with a detection accuracy improvement of 7.96% in 3D-mAP (IoU=0.5) and 2.78% in BEV-mAP (IoU=0.5) over the Multistage Fusion model, achieved the best performance among all feature fusion methods. Therefore, feature fusion of different modalities is affected by different strategies. Our solution utilizes cross-attention to effectively capture the correlation between heterogeneous modalities, whereas ViT leverages its powerful modeling capabilities to encode and reconstruct the fused features, thereby unifying the representation space of the two different modalities.

Finally, the ViT-FuseNet-V model shown in Table 2 refers to the case in which the proposed ViT-FuseNet framework uses only the vehicle point clouds as its LiDAR input. Without vehicle-infrastructure cooperation, a decrease in detection performance ranging from 0.27% to 1.82% is observed.

[1]The FFNet is originally proposed to address position misalignment and latency error issues in practical V2X communications. In this study, the data transmission from the infrastructure to the vehicle is assumed to be sufficiently good.

TABLE 4. Achievable performance with different position embedding methods.

Model	Positional embedding	$\text{Car}_{3D-mAP(IoU=0.7)}$			$\text{Pedestrain}_{3D-mAP(IoU=0.5)}$			$\text{Cyclist}_{3D-mAP(IoU=0.5)}$		
		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Ours-v1	No Pos.	73.95	66.42	60.64	49.04	38.46	38.03	46.69	42.27	41.70
Ours-v2	1d Pos.	**75.94**	**68.20**	**60.92**	**52.19**	40.35	**40.15**	**48.83**	42.77	**42.39**
Ours-v3	2d Pos.	74.45	67.29	60.68	50.55	**40.43**	40.09	45.68	**42.91**	42.28

This suggests that leveraging infrastructure information can lead to better performance than single-vehicle perception, mitigating the problem of detection bias. More importantly, it can enhance the system robustness.

In conclusion, with sufficiently strong V2X communication, our proposed method outperforms existing model perception methods. Moreover, as will be shown in Section IV-B2, the method introduced in this paper is applicable to various scenarios, including both single-vehicle and vehicle-infrastructure cooperation scenarios, showcasing a certain level of transferability.

2) EFFECT OF VIT MULTIMODAL FUSION MODULE

We further investigate the advantages of the ViT multimodal fusion module. The DAIR-V2X-C dataset only has labels for a single category of ''Car'' when cooperation, and the cooperative methods are only verified on ''Car''. To better illustrate the detection capability of the proposed method, we use the DAIR-V2X-V dataset, which provides three object categories, to conduct experiments. The four baseline models provided by [5] are considered as benchmark methods. The evaluation is carried out using the KITTI [32] evaluation method, which includes three levels of difficulty: easy, moderate, and hard. The experimental results are presented in Table 3. Clearly, our method consistently achieves better performance in terms of mAP compared with all other approaches. For instance, considering the best method among the baselines, that is, MVXNet [17], our method demonstrates an improvement of 3.78% in 3D-mAP (IoU=0.5) for ''vehicle'', 6.48% for ''pedestrian'', and 6.5% for ''bicycle'' in the moderate-level category.

It is observed from the results presented in Table 3 that multimodal fusion perception achieves superior performance improvements in detecting small objects, such as pedestrians and bicycles, compared with large objects. This is due to the fact that the information contained in point clouds is limited and includes only shape and depth. For small objects, the effective points in the point cloud are limited and sparse. Therefore, it is difficult to make accurate inference. For instance, objects such as street light poles and signs frequently exhibit point cloud features similar to those of pedestrians, resulting in false detections. The integration of multimodal features can effectively resolve this issue by utilizing the abundant texture and color information available in images. The effective implementation of multimodal feature fusion is in general a challenging task. Our framework provides a proper colution. In our study, the powerful cross-modal feature fusion capability of the cross-attention mechanism and the modeling capability of the ViT framework are employed to effectively integrate the features of images and point clouds into a common semantic space, resulting in a significant enhancement in multimodal perception performance.

C. IMPACT OF POSITION EMBEDDING

To evaluate the influence of different position embedding methods on the performance of the ViT multimodal fusion module, we conducted experiments with different methods of encoding spatial information, including no position embedding, 1D position embedding, and 2D position embedding. All the models undergo 50 epochs of training and are evaluated using the vehicle-side data of the DAIR-V2X-C dataset.

The achievable performance of the ViT FuseNet model is summarized in Table 4. It is evident that there is a performance gap between cases with and without position embedding. However, a small difference in terms of detection accuracy exists for the different methods of encoding positional information. We conjecture that encoding spatial information is less critical in our Transformer encoder because it operates on patch-level inputs instead of pixel-level inputs. The former has significantly smaller spatial dimensions than the latter. Consequently, learning to depict spatial relationships in this resolution is equally feasible for different position embedding strategies. Reference [25] also mentioned that position embedding is necessary when the vision Transformer is used to carry out object detection tasks, but position embedding methods of different dimensions have little impact on the results, because ViT perceives patch-level features.

V. CONCLUSION

In this study, we propose a novel multimodal feature fusion framework to conduct vehicle-infrastructure cooperative perception. The key component is a vision Transformer module designed based on a cross-attention mechanism, which effectively overcomes the challenge of integrating multimodal sensing data with distinct representation forms. Extensive experiments demonstrate that ViT-FuseNet notably enhances cooperative perception accuracy and outperforms existing advanced methods when data transmission is supported by high-quality V2X communication. Furthermore, our multimodal feature fusion module demonstrated effective fusion between different modalities compared with traditional fusion methods. It effectively captures the interrelationships between heterogeneous modalities, thereby

facilitating the efficient fusion of multimodal features. In addition, the framework exhibits a straightforward extension to single-vehicle perception tasks, showing a certain level of scalability.

Minimizing transmission costs, practical implementation and deployment are also of great importance in VIC3D tasks, and will be treated as a meaningful direction and studied in our future work.

REFERENCES

[1] C. Innes and S. Ramamoorthy, "Testing rare downstream safety violations via upstream adaptive sampling of perception error models," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 12744–12750.

[2] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019.

[3] W. Tong, A. Hussain, W. X. Bo, and S. Maharjan, "Artificial intelligence for vehicle-to-everything: A survey," *IEEE Access*, vol. 7, pp. 10823–10843, 2019.

[4] 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.*, vol. 35, 2022, pp. 4874–4886.

[5] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21329–21338.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[7] 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. Comput. Vis. (ECCV) 16th Eur. Conf.*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 605–621.

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

[9] D. Rukhovich, A. Vorontsova, and A. Konushin, "ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3D object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1265–1274.

[10] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. Eur. Conf. Comput. Vis.* Switzerland: Springer, 2022, pp. 1–18.

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

[12] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

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

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

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

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

[17] V. A. Sindagi, Y. Zhou, and O. Tuzel, "MVX-net: Multimodal Voxelnet for 3D object detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 7276–7282.

[18] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534.

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

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

[21] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2022, pp. 2583–2589.

[22] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.

[23] 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. Comput. Vis. ECCV 17th Eur. Conf.*, Tel Aviv, Israel. Cham, Switzerland: Springer, Oct. 2022, pp. 107–124.

[24] H. Yu, Y. Tang, E. Xie, J. Mao, J. Yuan, P. Luo, and Z. Nie, "Vehicle-infrastructure cooperative 3D object detection via feature flow prediction," 2023, *arXiv:2303.10552*.

[25] 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 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[27] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12124.

[28] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-SCNN: Fast semantic segmentation network," 2019, *arXiv:1902.04502*.

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Comput. Vis. ECCV 14th Eur. Conf.*, Amsterdam, The Netherlands. Cham, Switzerland: Springer, Oct. 2016, pp. 21–37.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[31] Z. Zhou, X. Zhao, Y. Wang, P. Wang, and H. Foroosh, "CenterFormer: Center-based transformer for 3D object detection," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 496–513.

[32] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

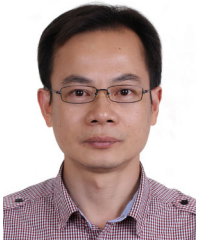
[33] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, Nov. 2019, pp. 88–100.

[34] H. Yu, Y. Zhao, Y. Zou, Q. Li, H. Yu, and Y. Ren, "Multistage fusion approach of LiDAR and camera for vehicle-infrastructure cooperative object detection," in *Proc. 5th World Conf. Mech. Eng. Intell. Manuf. (WCMEIM)*, Nov. 2022, pp. 811–816.

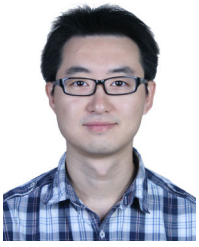
[35] Y. Lu, Q. Li, B. Liu, M. Dianati, C. Feng, S. Chen, and Y. Wang, "Robust collaborative 3D object detection in presence of pose errors," 2022, *arXiv:2211.07214*.

YANG ZHOU received the B.E. degree in communication engineering from Tongji University, Shanghai, China, in 2021, where he is currently pursuing the M.E. degree with the College of Electronic and Information Engineering. His research interests include deep learning-based multimodal fusion and vision transformer in vehicle-infrastructure cooperative perception.

CAI YANG received the B.E. degree in communication engineering from Tongji University, Shanghai, China, in 2023, where she is currently pursuing the M.E. degree with the College of Electronic and Information Engineering. Her research interests include multimodal object detection for fusion of vision and LiDAR in computer vision.



PING WANG (Member, IEEE) received the Ph.D. degree in computer science from Shanghai Jiaotong University, Shanghai, China, in 2007. He is an Associate Professor of information and communication engineering with Tongji University. He has published over 120 scientific articles and three books. He has applied one international patent and over 20 national patents. His main research interests include routing algorithms, resource allocation of wireless networks (especially for VANETs), and multi-sensor information fusion. He also focuses on research, development, and verification of connected intelligent vehicles.



CHAO WANG (Member, IEEE) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2003, and the M.Sc. and Ph.D. degrees from The University of Edinburgh, Edinburgh, U.K., in 2005 and 2009, respectively. He is a Professor with Tongji University, Shanghai, China. In 2008, he was a Visiting Student Research Collaborator with Princeton University, Princeton, USA. From 2009 to 2012, he was a Postdoctoral Research Associate with the KTH-Royal Institute of Technology, Stockholm, Sweden. From 2018 to 2020, he was a Marie Curie Fellow with the University of Exeter, Exeter, U.K. His research interests include information theory and signal processing for wireless communication networks, data-driven research and applications for smart city, and intelligent transportation systems.

XINHONG WANG received the Ph.D. degree in computer application technology from Northeastern University. She is an associate professor. In recent years, she has published over 30 academic articles and applied over ten national patents. Her research focuses on the Internet of Vehicles and wireless communication networks.

NGUYEN NGOC VAN received the bachelor's and master's degrees in electronics and telecommunications (major) from Hanoi University of Science and Technology, in 2000 and 2003, respectively, and the Ph.D. degree from Tongji University, in 2012. He received the Vietnam Government Scholarship for the Ph.D. study. He was with Tongji University as a Ph.D. Postgraduate. He is a Lecturer with the School of Electrical and Electronic Engineering, Hanoi University of Science and Technology. Based on the research and development of broadband wireless networks (5G/6G) and the Internet of Vehicles (IoV) networks, he has gone to in-depth research on technology solutions to optimize and propose network-based applications of the IoV and C-V2X.