BEV-V2X: Cooperative Birds-Eye-View Fusion and Grid Occupancy Prediction via V2X-Based Data Sharing

Cheng Chang, Jiawei Zhang, Kunpeng Zhang, Wenqin Zhong, Xinyu Peng, Shen Li, Li Li, *Fellow*, *IEEE*

***Abstract*— Birds-Eye-View (BEV) perception can naturally represent natural scenes, which is conducive to multimodal data processing and fusion. BEV data contain rich semantics and integrate the information of driving scenes, which play an important role in researches related to autonomous driving. However, BEV constructed by single vehicle perception encounter certain issues, such as low accuracy and insufficient range, and thus cannot be well applied to scenario understanding and driving situation prediction. To address the challenges, this paper proposes a novel data-driven approach based on vehicle-to-everything (V2X) communication. The roadside unit or cloud center collects local BEV data from all connected and automated vehicles (CAVs) within the control area, then fuses and predicts the future global BEV occupancy grid map. It provides powerful support for driving safety warning, cooperative driving planning, cooperative traffic control and other applications. More precisely, we develop an attention-based cooperative BEV fusion and prediction model called BEV-V2X. We also compare the performance of BEV-V2X with that of single vehicle prediction. Experimental results demonstrate that our proposed method achieves higher accuracy. Even in cases where not all vehicles are CAVs, the model can still comprehensively estimate and predict global spatiotemporal changes. We also discuss the impact of the CAV rate, single vehicle perception ability, and grid size on the fusion and prediction results.**

***Index Terms*—Cooperative Driving, Birds-Eye-View Fusion, Occupancy Prediction**

I. INTRODUCTION

INTELLIGENT vehicles rely on accurate environment awareness to achieve robust and safe autonomous driving [1]-[2]. Driving environment includes static elements such as road layout and lane markings, as well as dynamic elements such as vehicles and pedestrians [3]. Classical perception systems of intelligent vehicles collect driving environment data [4]-[6] through sensors such as cameras, LiDAR and Radar, and perform tasks such as image classification, scene segmentation and object detection to understand driving scenes. To obtain more accurate results, multi-sensor data fusion methods [7]-[9] are widely used by researchers.

Perception results have various representation formats. Recently, the Birds-Eye-View (BEV) format has recently attracted the attention of many researchers. Many researches propose the methods that aggregate the raw sensory data, such as camera images and Lidar points, from the perspective view to the birds-eye-view on the vehicle side [10]-[11]. BEV clearly represents the position and state of scenario objects in compact and semantic way based on occupancy grid [12]-[13]. BEV is usually represented as an image with a certain resolution. Each pixel corresponds to a certain size of grid unit (e.g., 1 pixel corresponds to $0.5m$ square grid).

BEV data contain rich semantics, which can integrate environment information, road information, and static and dynamic information of traffic participants. As the processing and enhancement of raw sensory data, BEV data are conducive to fusion in unified space. Single frame of BEV can describe the driving state and relationship of each traffic participant at a certain time stamp in certain road environment, which is consistent with the concept of scene. Continuous frames of BEV can represent the driving behavior and interaction of traffic participants within a certain spatial and temporal range, which is consistent with the concept of scenario [14]. BEV data can play an important role in systematic and unified scenario representation in research fields such as motion prediction [15]-[17], trajectory planning [18]-[21], and intelligence testing [22]-[24].

However, the BEV constructed by single vehicle perception may have some problems, such as low accuracy and insufficient range. Fig. 1 shows the BEV perception centered on vehicle A, B and D respectively in an intersection scenario. If the roadside unit is not deployed in the scenario, due to the restrictions of road geometry and limited perception range of single vehicle, vehicle A, vehicle B, and vehicle D cannot directly achieve mutual perception. It leads to poor understandings of the global scenario. If vehicle A chooses to turn right at the intersection, it may cause safety risks due to insufficient warning time advance. In addition, vehicle C is lo-

This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFB2501200, and the National Natural Science Foundation of China under Grant 52272420. (Corresponding author: *Shen Li, Li Li*)

Cheng Chang, Jiawei Zhang and Xinyu Peng are with the Department of Automation, Tsinghua University, Beijing 100084, China.

Kunpeng Zhang is with the Department of Automation, BNRist, Tsinghua University, Beijing 100084, China and the College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China.

Wenqin Zhong is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518057, China.

Shen Li is with the Department of Civil Engineering, Tsinghua University, Beijing 100084, China (e-mail: sli299@tsinghua.edu.cn).

Li Li is with the Department of Automation, BNRist, Tsinghua University, Beijing 100084, China (e-mail: li-li@tsinghua.edu.cn).

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

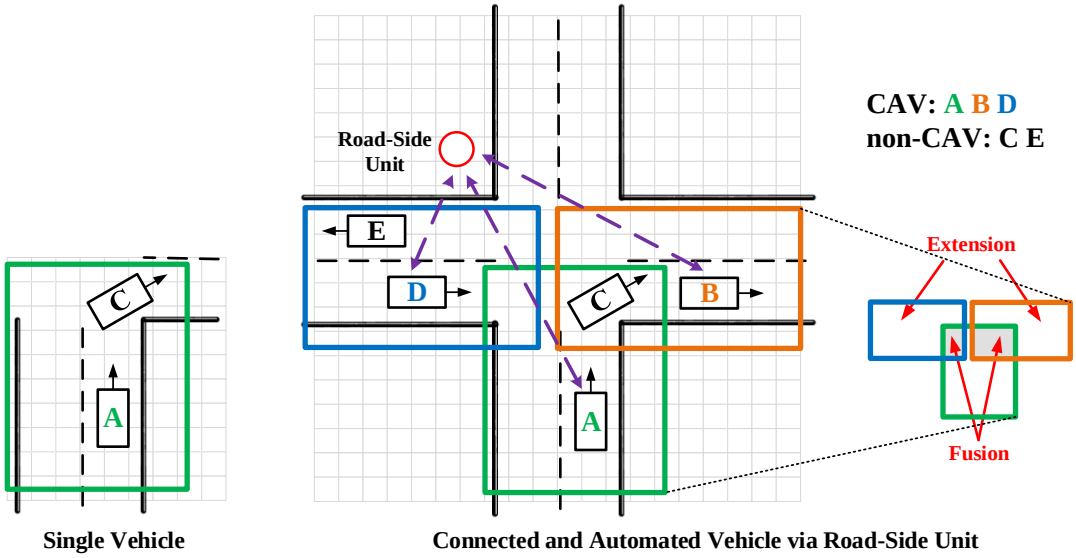


Fig. 1. An illustration of BEV perception based on single vehicle and CAV relative to vehicle A in an intersection scenario.

cated in the overlapping perception space of vehicle A and B. After raw perception data are aggregated to the BEV, the corresponding grid position and associated confidence of vehicle C in the local BEV of A and B are also different. It is a critical issue to collect and fuse the local BEV of different vehicles to obtain global BEV with higher reliability and more comprehensive scenario understanding. Through the fusion information, we can also make inferences and predictions for future driving situations. It can provide more accurate and comprehensive information for downstream tasks, such as driving safety warning, trajectory planning, and traffic control.

Vehicle-to-everything (V2X) communication provides a new approach to solve the problem of data fusion for different vehicles and infrastructures. In cooperative driving scenario, each vehicle is tightly linked with other vehicles or roadside units [25]-[27]. Each connected and automated vehicle (CAV) regularly reports its own information to other vehicles or roadside units. By aggregating and fusing the data information from different CAVs, we can get a more accurate understanding of the global scenario. As shown in Fig. 1, we show the differences between single vehicle perception and V2X cooperative perception. The perception range of single vehicle is limited to the local area. With V2X communication, the area that can be perceived by vehicles can be significantly extended. Moreover, via the BEV perception of CAVs, we can partially recover and obtain the information of non-CAVs. Multi-source information fusion can also be conducted on the overlapping BEV space among vehicles to improve the perception accuracy.

In this paper, we propose a BEV fusion and prediction model BEV-V2X based on V2X communication and attention neural network model in cooperative driving scenario. The roadside unit or cloud center can collect the local BEV data of all CAVs in the control area. By extracting the single vehicle BEV data in the historical time horizons, we can integrate the perception information of different CAVs and predict the global BEV occupancy grid map in the future time horizons. This paper focuses on BEV fusion and prediction. The fusion and prediction results can help achieve accurate environment perception and strengthen the understanding of the global scenario. Based on the results, the system can provide real-time driving risk warning, formulate the corresponding planning scheme, and send the messages to vehicles in the control area. It provides powerful support for driving safety warning, cooperative driving planning, cooperative traffic control and other applications.

In fact, there are basically two modes of vehicle-to-vehicle (V2V) communication and vehicle-to-infrastructure (V2I) communication to achieve BEV fusion via V2X technique. We recommend that the corresponding models be deployed at the roadside or cloud center. The following models are also explained based on this basis, and the specific advantages are further discussed in *Section V*.

Our main contributions are as follows:

1) We propose a new BEV fusion and occupancy prediction method via V2X communication. Experimental results show the advantages of attention model, and the global cooperative BEV fusion and prediction results achieve high accuracy.

2) We compare the prediction method of single vehicle based on its own local BEV and the prediction method of BEV fusion via V2V communication. Experimental results show that the BEV fusion and prediction method based on edge computing/cloud computing via V2I communication is much better.

3) When the cooperative driving scenario is not full of CAVs, our proposed method can still obtain the partial recovery information of non-CAVs and the global scenarios well. We also show that as the CAV rate of the control area increases, we can obtain richer information for fusion and get more accurate global cooperative BEV results. As the perception ability of single vehicle becomes stronger, the uncertainty of single BEV results will decrease, and the fusion and prediction results will also be better.

The rest of the paper are arranged as follows: *Section II*

reviews the related works. *Section III* introduces the data flow and the specific neural network model structure. *Section IV* conducts simulation experiments on naturalistic datasets, and experimental results demonstrate the effectiveness of the proposed approach. *Section V* discusses the advantages of V2X based fusion and prediction compared to single vehicle prediction. *Section VI* summarizes the paper.

II. RELATED WORKS

A. BEV Perception and Prediction

BEV perception on the vehicle side mainly focuses on the data conversion from perspective view to birds-eye-view. The perception module mainly takes the raw sensory data as input, such as camera images and Lidar point clouds, and generates the occupancy grid map under birds' eye view as output. Some researchers use inverse perspective mapping (IPM) methods [28] to learn the transformation matrix. Recent works use deep learning-based data-driven models, such as multi-layer perceptron (MLP) [29], convolution neural network (CNN) [12], [30] and attention-based models [13], to transform the data. The above methods can aggregate the spatiotemporal features from multiple sensors and fuse them into a unified BEV space.

BEV data can represent driving scenarios and help to effectively extract the states and interactions of the traffic participants. The data are frequently used as input in motion prediction related modules [31]-[32]. According to the output format, BEV based motion prediction researches can be categorized into two types, BEV-based trajectory prediction and BEV occupancy prediction. BEV-based trajectory prediction task is to predict the candidate trajectories of each vehicle and corresponding confidences in future time horizon with BEV as input [15], [33]. While BEV occupancy prediction task is defined as predicting the occupancy grid state within a certain spatial area in future time horizons, which has consistency with trajectory prediction [34]. The output BEV occupancy grid map results enable the system to directly obtain the global driving situation without processing the vehicles' trajectories separately. Many researches also perform joint perception and prediction in multi-task modal using unified BEV representations [35]-[36].

B. V2X-based Data Fusion

V2X communication technology enables vehicles and infrastructures to share their data for information fusion. We review the related works in terms of different data types to be transmitted.

Trajectory data. In many previous works related to cooperative driving based on V2X communication, it is assumed that all vehicles are CAVs [37]-[38]. Therefore, vehicles can directly send position and trajectory data to other vehicles or roadside units, and the global scenarios and states can be extracted after data aggregation. Trajectory data can help to avoid vehicle perception failure, which also occupy less storage space, require less computation, and facilitate data transmission. Many researches address the location and state estimation of multiple vehicles via trajectory data exchange. The methods include multi-source confidence weighting [39]-[40], Kalman filter [41], extended Kalman filter [42], etc. The methods use the motion equations for observation and prediction, and gradually update the global localization of the vehicles. However, it is difficult to cope with the complex nonlinear situation. Moreover, in real cooperative driving scenario, it is common that not all vehicles are CAVs [43]-[44]. Relying only on motion or trajectory data may not work in these cases.

Raw sensory data. Some researches propose the method of aggregating the raw sensory data from CAVs and fusing the information to promote the perception. [45] propose Cooper system to fuse the 3D Lidar point clouds collected from different positions and angles of CAVs. [46] propose Autocast to fuse Lidar data according to the visibility and relevance of vehicles. [47] and [48] combine the data from vehicles and roadside units to perform the Lidar-based data fusion and object detection. The above methods realize the raw level data fusion using rich perceptual information. However, the raw sensory data occupy a large amount of storage space and reduce the speed of data transmission and interaction, thus cannot ensure the real-time performance of autonomous driving systems.

Intermediate feature data. To maintain the balance between perception accuracy and data transmission delay, some researches focus on the intermediate feature [49] based data fusion methods. The intermediate features are mainly generated by deep learning-based encoders deployed on vehicles, such as CNN [50]-[51], and attention-based models [52]-[54]. The aggregated features are further decoded on the vehicle side or road side to generate the final perception results. The decoders mainly include CNN [52], graph-based models [51], [55], and attention-base models [53]. The overall solutions of the above methods are mainly tightly coupled end-to-end models. As the compressed representation of raw sensory data, the intermediate feature data can reflect the perception information to a certain extent with the reduction of communication bandwidth. However, the formats of the intermediate features are not unified in many researches. In addition, the generated features are always semantically unexplained, which reduces the interpretability for humans and the reusability for other tasks.

Compared with the above methods, our V2X-based BEV fusion and prediction approach has the following differences:

1) Compared to trajectory data, BEV data contain more perception information and extend more perception space. With the help of BEV data, we can partially recover the state of non-CAVs and obtain more accurate global results through the spatial and temporal data fusion in the unified BEV space.

2) Compared to raw sensory data, BEV data are the aggregation and summarization of the vehicle's raw sensory information. The BEV format can reduce memory storage and save transmission bandwidth. According to the points generated by 64-laser Lidar per second, the size of raw data is up to 150 Mbit. In contrast, the size of BEV data package is less than 1Mbit. The data transmission delay of BEV data can

guarantee the real-time performance according to the current development of V2X communication protocol [56]. We will introduce the data structure in detail in *Section III* and *Section IV*.

3) Compared to intermediate feature data, the proposed method has the following differences. First, the explainable BEV data are generated on the vehicle side and transmitted to the road side for fusion and prediction. The methods can fully utilize the perception ability of single vehicles, and construct the loosely coupled hierarchical model. Second, as a typical representation format of driving scenario, BEV data contain rich scenario semantics and human-machine understanding. The use of scenario data can increase the interpretability and trust of data-driven models [57]. Third, BEV data are also convenient and useful for many other modules related to autonomous driving, such as trajectory planning, motion control, and intelligence testing. The reusability of data can be significantly increased by transmitting BEV instead of intermediate features.

4) Compared to the upstream perception tasks, our focus is on obtaining global BEV fusion and prediction results and supporting the downstream tasks, such as driving safety warning, cooperative driving planning, and cooperative traffic control. Although the data metrics cannot be directly compared due to different data sources and tasks, according to the results of real-time scene segmentation and object detection related to BEV perception with high accuracy, the performance of our proposed approach is acceptable and valuable. The details will be further illustrated in *Section IV*.

III. A DATA DRIVEN MODEL

A. The Data

Fig. 2 shows the overall data flow of the fusion and prediction process. The roadside unit collects the local BEV of all CAVs in the control area, and periodically extracts the historical data. Using the data, the system calls the deep learning model to fuse the information from different vehicles, and predict the future cooperative BEV occupancy grid map.

The data and phrase definitions of each side in the framework are described in detail as follows:

1) Vehicle side

With the help of various sensors, such as cameras and Lidar, the single vehicle perceives the surrounding environment. Then, the vehicle system converts the raw sensory data, such as images and point clouds into BEV space, and generates the local BEV centered on its own coordinates. BEV is a semantically composite data structure, which uses matrices to represent the occupancy of scenario elements within a certain spatial area. Each matrix element corresponds to the occupancy probability or state of each grid in the driving environment, which can be further summarized and displayed as RGB image.

Single Vehicle BEV Probability (SBEV-P)

At each grid location of the BEV, the occupying objects may include both vehicles and road elements, and they are not in conflict with each other. Therefore, we divide the scenario elements into different categories, i.e., dynamic traffic participants such as vehicles and pedestrians, and static road environment information such as drivable areas, lanes, traffic infrastructures, channelization, etc. We apply the symbol $P_c(x, y)$ to denote the probability that the BEV position (x, y) is occupied by category c. $[P_c(x, y)]_{C\times H\times W}$ is the occupancy probability matrix of C elements in the grid network with the size of $H \times W$.

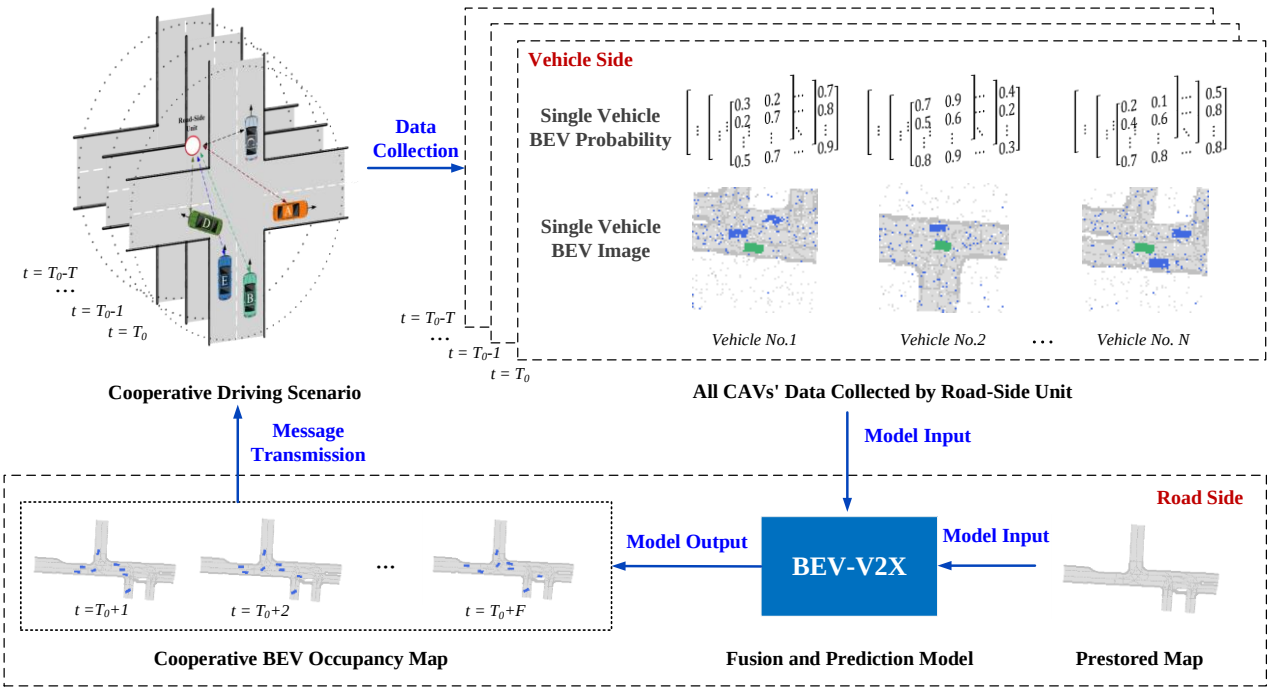


Fig. 2. The data flowchart of cooperative BEV fusion and prediction.

Single Vehicle BEV Image (SBEV-I)

Each BEV grid has a corresponding probability estimate for different categories. By summarizing the occupancy probability representation into occupancy state representation, the corresponding BEV RGB image can be further displayed. We use $M_c(x,y)=\{0,1\}$ random variable to represent whether grid position (x,y) is occupied by category c. By setting certain threshold *TH*, the transformation from probability representation to state representation is shown in equation (1).

$$M_c(x,y)=\begin{cases}1, & P_c(x,y)>TH\\ 0, & P_c(x,y)<TH\end{cases} \quad TH\in(0,1) \tag{1}$$

where *TH* is usually set as 0.5. $[M_c(x,y)]_{C\times H\times W}$ is the occupancy state matrix of C elements in the grid network with the size of $H\times W$.

We assign different RGB values to each category and set $Priority_c$ for all categories in the category set S_c. For example, the priority of vehicles and pedestrians is higher than that of road information, and the priority of traffic infrastructures is higher than that of lanes. In this way, the final color assigned to each grid position (x,y) in the BEV image follows equation (2).

$$RGB(x,y)=RGB(\underset{\{c\in S_c|M_c(x,y)=1\}}{argmax}\ Priority_c) \tag{2}$$

By further distinguishing the colors of the current ego vehicle from the perceived surrounding vehicles (such as green and blue in Fig. 2), the BEV image centered on the ego vehicle can be obtained. The tensor size of SBEV-I is $3\times H\times W$.

SBEV-P represents the estimated probability information of the driving environment with the help of vehicle perception ability. While SBEV-I implies the transformation rules and contain richer semantics via the adjacent image pixels, which are close to human understanding. The combination of SBEV-P and SBEV-I can comprehensively reflect the vehicle perception results and will also be applied to the input of our proposed model.

CAV sends its real-time SBEV data package to the roadside unit in a timer-trigger style [58]. In *Section IV*, we will explain the impact of the SBEV grid size on transmission, collection and storage with specific naturalistic data.

2) Road side

The roadside unit collects the local BEV information of all CAVs in the control area, and periodically extracts data in the historical time horizon. Combined with the internal pre-stored grid map data of the control area, the system calls the deployed BEV-V2X fusion and prediction model, and obtains the cooperative BEV occupancy of the global scenario in the future.

PreStored Map (PMap)

The roadside unit can store the map representation of the control area in advance. We also use $M_c(x,y)=\{0,1\}$ random variable to indicate whether the grid position (x,y) in the map is occupied by category c. Differently, here the categories include all static elements except dynamic traffic participants, such as drivable area, lanes, traffic infrastructures, etc., that is $c\in S_c\setminus Dynamic_c$. Assuming that there exist static elements of MC classes, the $[M_c(x,y)]_{MC\times HO\times WO}$ matrix is the occupancy state representation of MC elements in the grid map with the size of $HO\times WO$.

Cooperative BEV Fusion and Prediction Model (BEV-V2X)

The system extracts SBEV data in historical time horizon T, and builds the fusion and prediction model BEV-V2X based on spatiotemporal attention. The specific model structure will be detailed in *Section III.B*.

Cooperative BEV Occupancy Map (CBEV)

The model outputs the BEV occupancy grid estimate for the whole control area in future time horizon F. The output $[P_c(x,y)]_{C\times HO\times WO}$ is the occupancy probability of C elements in the global grid network with the size of $HO\times WO$. Further, the occupancy state representation and visual image of the global CBEV are generated by the transformation rules in formula (1) and (2), which constitute the final fusion and prediction information.

B. The Structure of BEV-V2X

In this paper, the BEV fusion and prediction model of cooperative driving scenario emphasizes the spatial and temporal interaction of CAVs within the control area of roadside units. Therefore, we use the attention mechanism of the transformer model [59] to capture the spatiotemporal relationship of the data. The model outputs the cooperative BEV occupancy prediction of the global scene. Previous studies [33], [60] have illustrated the effectiveness of attention models in prediction tasks.

The overall BEV-V2X model is shown in Fig. 3, with the SBEV obtained by each CAV perception as the external input, i.e., the normalized connection tensor of SBEV-I and SBEV-P in *Section III.A*, and the pre-stored map of the control area as the internal input, that is, the PMap in *Section III.A*. The model outputs the future cooperative BEV occupancy grid map. In addition, the model also estimates its global road map representation to accelerate the training convergence.

The BEV-V2X model can be roughly divided into three parts: data collection and embedding based on convolution model, fusion and prediction of spatiotemporal data features based on attention model, and upsample and construction of the global BEV based on deconvolution.

In the first part, the dimensions of SBEV-I and SBEV-P are $(3,H,W)$ and (C,H,W) respectively. The total dimension of the perception tensor is $(B,T,N,C+3,H,W)$, where B is the batchsize of the model, T is the historical time horizon, N is the maximum number of possible vehicles in the control area.

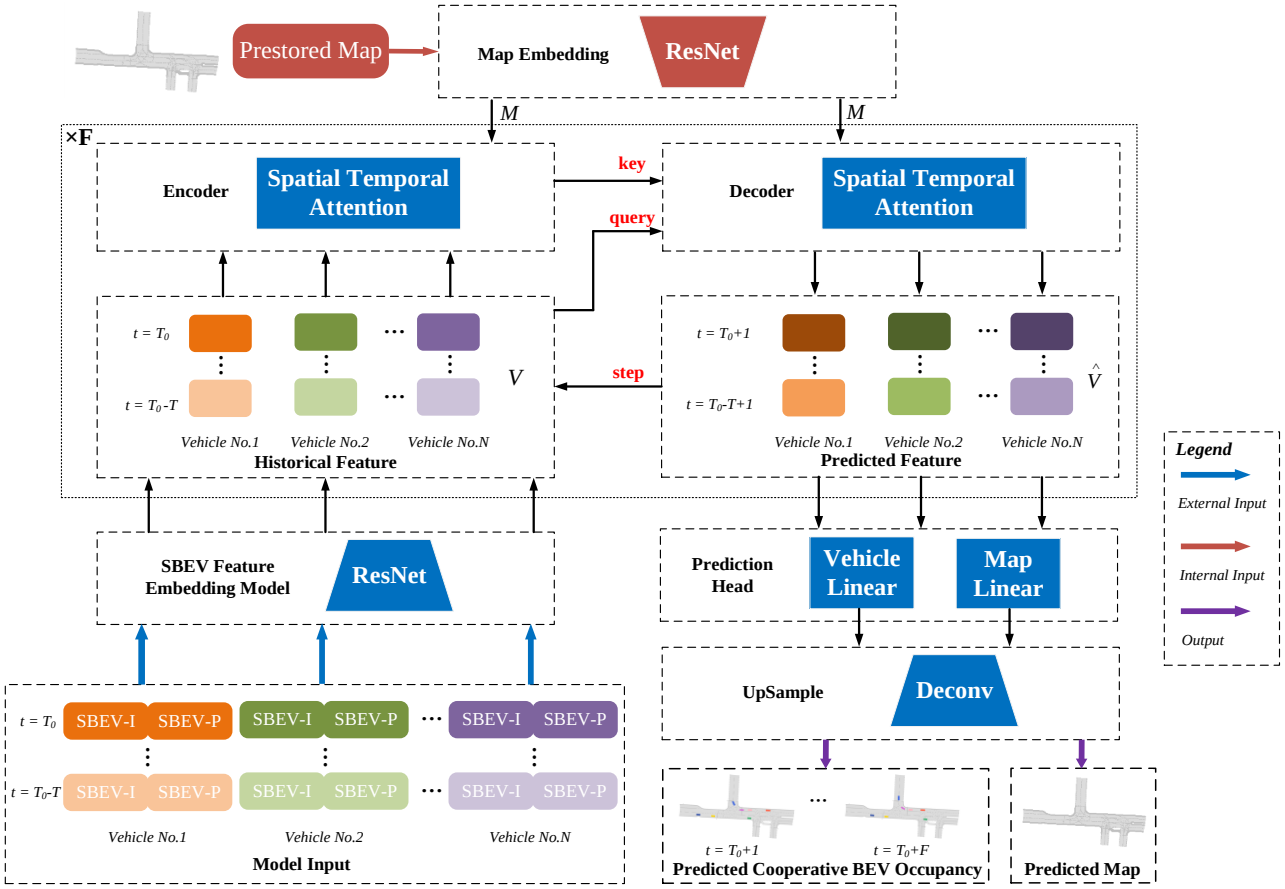


Fig. 3. The structure of BEV-V2X.

Since the number of CAVs' data packages received by roadside unit at each timestamp may be uncertain, to maintain the size unity during model calculation, we choose the upper limit of the possible CAV number as N. The model marks the valid CAV data when receiving data stream, and the rest will be processed by corresponding padding. The dimension of the original map tensor pre-stored by the roadside unit is (B, MC, HO, WO).

We apply Resnet [61] to embed the input SBEV data, and obtain the tensor V with dimension (B, T, N, E), where E is the embedding dimension of each SBEV data group. The tensor will be used for subsequent spatiotemporal attention extraction and prediction. The similar Resnet model is applied for map data embedding, and the tensor M with dimension $(B, 1, N, E)$ is obtained.

In the second part, we use the spatiotemporal attention model to extract the spatiotemporal interactions of CAVs' SBEV data in the control area.

Attention mechanism of transformer model is initially applied in the NLP field, and is used to conduct machine translation via encoder-decoder architecture. Here we follow the similar multi-head attention approach. For each attention module, the (q, k, v) triplet input should be conducted scaled dot-product operation and softmax normalization operation. The number of attention heads is set as h, and the projections of each attention head are $Q_i \in R^{d_q}, K_i \in R^{d_k}, V_i \in R^{d_v}$. The embedding dimensions of the three tensors in this paper are the same, which are set as $E = d$. Thus, the embedding dimension of each attention head follows:

$$d_q = d_k = d_v = \frac{d}{h} \tag{3}$$

The parameters of the model include h groups of different weight matrix pairs (W_i^q, W_i^k, W_i^v), whose dimensions correspond to $(d \times d_q, d \times d_k, d \times d_v)$. The obtained query, key and value are:

$$Q_i = qW_i^q, K_i = kW_i^k, V_i = vW_i^v \tag{4}$$

The calculated attention tensor is:

$$head_i = Attention(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{5}$$

By concatenating the results of each head, the final attention of multi-heads can be obtained as follows:

$$MultiHead = Concat\ (head_i, ..., head_h)W^O \tag{6}$$

where the dimension of W^O is $d \times d$.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

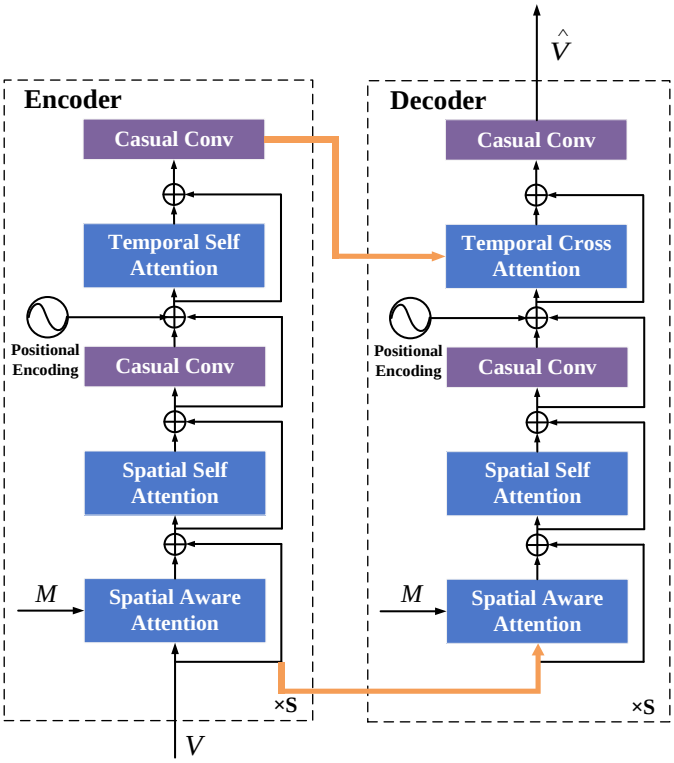


Fig. 4. The structure of spatial temporal attention.

The spatiotemporal attention model is shown in Fig. 4, which adopts encoder-decoder architecture. The dimension of the input V tensor is (B,T,N,E). First, the global spatial aware attention is conducted with the roadside unit pre-stored map, where the map embedding M is expanded to (B,T,N,E) along the time axis. In this step, attention related calculation is operated at the spatial level. We transpose and reshape V and M tensors to the dimension of $(B\times T,N,E)$. We take V as query, M as key and value for the attention operation, the dimension of output tensor is $(B\times T,N,E)$ and is further reshaped as (B,T,N,E). The global spatial aware attention is equivalent to adding global spatial position encoding to the original input data, so that we can fuse CAVs' positions in cooperative BEV and be aware of the coordinate system transformation.

Next, we conduct the spatial self attention operation. The input data are taken as query, key and value simultaneously, and are conducted at the spatial level similar to previous operations. The spatial self-attention is operated to be aware of vehicle's spatial relations, which facilitates the fusion of CAVs' BEV perception results.

Then, we extract the temporal interaction of the input data. We add a learnable temporal position encoding to fully represent the time series position of each CAV's data. In the encoder module, we reshape the dimension of tensor to $(B\times N,T,E)$ and conduct the temporal self attention operation. The dimension of output tensor is $(B\times N,T,E)$ and is further reshaped as (B,T,N,E). The prior physical information of vehicle motion determines that there exist temporal relations between the BEV data of continuous frames. Here we adopt temporal cross attention in the decoder module, which makes the decoder query interact with the encoder outputs to accurately capture the temporal evolutions. The design can help the model make better inferences and predictions along the time axis.

Finally, in the encoder and decoder of spatiotemporal attention, we further add casual convolution with dilation. Causal and dilated convolution [62]-[64] shows high performance in time series prediction task, which can expand the receptive field without increasing parameters. Thus, the top layer can use a wider range of information in the input data layer through the dilation. We reshape the dimension of tensor to $(B\times N,E,T)$ and conduct the casual convolution operation. The time length of the output satisfies $T_{out}=T$.

In the third part, the prediction feature of the future time horizon F, whose dimension is (B,F,N,E), is conducted linear and upsample operation to output the predicted cooperative BEV occupancy. Due to the large size of cooperative BEV, the compressed feature of dimension $(B,F,HO/4,WO/4)$ is output through the vehicle linear layer, and the tensor of dimension $(B,MC,HO/4,WO/4)$ is output through the map linear layer. Then after two layers of deconvolution and upsample, the tensors are output as the predicted cooperative BEV occupancy and the predicted road map, which correspond to the dimensions of (B,F,HO,WO) and (B,MC,HO,WO) respectively.

Since the map information is pre-stored on the roadside, the system only needs to take the predicted grid occupancy state of vehicles and pedestrians, and then concatenate the tensors with the pre-stored standard map occupancy state as the final results.

We further consider the computation complexity of the model. The complexity of the convolution and embedding model in the first part is $O(HW+E)$, the complexity of the spatiotemporal attention model in the second part is $O(E^2+E)$, and the complexity of deconvolution and upsample in the third part is $O(E^2)$. Since the tensor dimension E of the embedded data is far smaller than the size HW of each source SBEV matrix. Using the embedded tensor for attention calculation can significantly reduce the complexity and computation time.

C. The Metrics and Loss Function

The output of cooperative BEV is the probability matrix of grids occupied by dynamic traffic participants at each future timestamp. The probability matrix element is denoted as P_t^s, where t represents the temporal position and s represents the spatial position. The grid occupancy state prediction M_t^s is estimated by the probability, which can be transformed according to equation (1). It is assumed that the grid occupancy state label is represented as O_t^s. We use Intersection over Union (IOU) to evaluate the effectiveness of the output CBEV as follows:

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

$$IOU_t(M,O)=\frac{\sum_{x,y} M_t^s O_t^s}{\sum_{x,y} M_t^s+\sum_{x,y} O_t^s-\sum_{x,y} M_t^s O_t^s} \tag{7}$$

By calculating the IOU between the global CBEV prediction and CBEV occupancy label, we can compare the prediction accuracy and further evaluate the effectiveness of the model. The higher the IOU index, the more parts of the prediction results coincide with the true label, and the model fusion and prediction accuracy is better.

During model training, in addition to the loss function related to IOU, we also add the loss related to binary cross entropy (BCE) to accelerate the convergence of the model and ensure that the foreground (the vehicle map) of the model will not be merged by the background (the road map). BCE loss function is expressed as follows:

$$BCE_t(P,O)=-\frac{\sum_{x,y} O_t^s \log P_t^s+(1-O_t^s)\log(1-P_t^s)}{HO\times WO} \tag{8}$$

The probability of predicted vehicle occupancy is denoted as $P_t^s(V)$, and the probability of predicted map occupancy is denoted as $P_t^s(M)$. The total loss function is as follows:

$$\begin{aligned} loss(P,O)=\lambda_V\sum_t IOU_t(P(V),O(V))+BCE_t(P(V),O(V))+\\ \lambda_M\sum_t IOU_t(P(M),O(M))+BCE_t(P(M),O(M)) \end{aligned} \tag{9}$$

where λ_V,λ_M are the coefficients of vehicle occupancy and road map occupancy.

IV. Testing Results

To verify the effectiveness of the method, we conduct a series of simulation experiments. First, we verify that the global BEV fusion and prediction information output by the model achieves high accuracy, which can achieve better results compared to other deep learning-based models. Second, when the vehicles in the scenario are not full of CAVs, we verify that the proposed method can still better obtain the recovery information of non-CAVs and global scenes. Finally, we observe the influence of the CAV rate, single vehicle perception ability, and grid size on the final BEV fusion and prediction results.

A. Experiment Setting

The simulation experiment of BEV fusion and prediction requires naturalistic driving scenario data, which contain the movement information of traffic participants in a certain spatial area and a continuous time range, as well as the environment information. Currently there are many scenario datasets, including NGSIM [65] and HighD [66] and other datasets are collected on highways, which are mostly straight roads. The driving behaviors are relatively trivial, and not typical for comparison. Argoverse [67] dataset does not provide specific size information of traffic participants, such as the length and width of the vehicle. Although it is suitable for trajectory prediction, it is not suitable for occupancy prediction, where vehicle size parameters need to be specifically considered. Interaction dataset [68] contains a variety of rich interactive driving scenarios that provide vehicle motion data and road map data, etc. The duration of each vehicle's trajectory is relatively long. Therefore, we focus on the model evaluation using the Interaction dataset. We use data framework [14] to preprocess and annotate the trajectory data, map data and drivable area.

We select the typical intersection scenario dataset to demonstrate our simulation. DR_USA_Intersection_EP0 consists of 8 data parts, each lasting about 6 minutes. Referring to the observation and prediction time parameters set by previous motion prediction works [15], [17], [33], and further considering the time statistics of the dataset [68], we set the historical time horizon to 3s, and the prediction time horizon to 3s. We divide the training set, validation set and test set according to the ratio of 8:1:1.

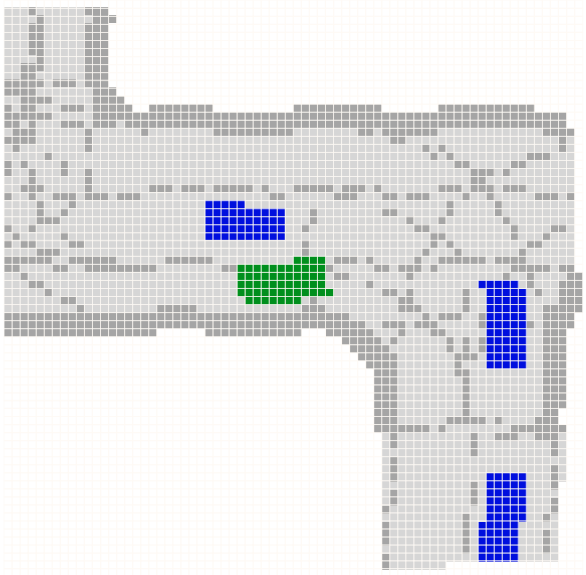


Fig. 5. An example of SBEV-I with 0.5m grid.

In the intersection scenario, the range of single vehicle perception is limited due to factors such as road geometry and vehicle occlusion, we set the SBEV range to a square area of 36m around each vehicle. In addition, we set the CBEV range to a square area of 144m, which is occupied by the categories of drivable area, road markings, and vehicles, respectively. Following previous work on autonomous driving perception and prediction modules, each pixel of the BEV image corresponds to a grid size of 0.5m [13], [15]. As shown in Fig. 5, we show the SBEV-I example for the 0.5m grid size. Due to the dense grid lines, we do not specifically show the grids in the following experimental results to make the BEV results more clear.

According to the above data parameters, the size of the data package transmitted by CAV is 0.62Mbit once the time is triggered. When the single vehicle perception range is from 15m to 50m, and the category C varies from 2 to 5, the size of

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

each package transmitted by CAV is from 0.08Mbit to 1.84Mbit. Under the communication protocol related to the internet of vehicles [56], [69], the transmission delay to the roadside can be kept below 70ms. The transmitted data of each CAV can be aggregated, stored and extracted through the CD-DB model [38] deployed on the roadside unit. According to experimental tests, the average time of data operations can be kept below 10ms. Under the computation complexity analyzed in *Section III.B*, the computation and inference time of the deep learning model can be below 50ms. Thus, the real-time capability of driving safety warning and motion control modules can be guaranteed.

In *Section III.A*, we denote the occupancy probability of BEV position (x, y) by C elements as $P_c(x, y)$, which is obtained from the perception module of the single vehicle. The level of the probability close to the true label can represent the perception ability of vehicles. The probability result SBEV-P constructed by different perception data sources will correspond to different distribution functions under different perception modules, among which the most typical one is Beta distribution. In [70]-[71], the grid occupancy of the map is set as a random variable satisfying the Beta distribution. The corresponding probability density function $x \sim Beta(\alpha, \beta)$ is as follows:

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad x \in [0,1] \tag{10}$$

where $B(\alpha, \beta)$ is as follows:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{11}$$

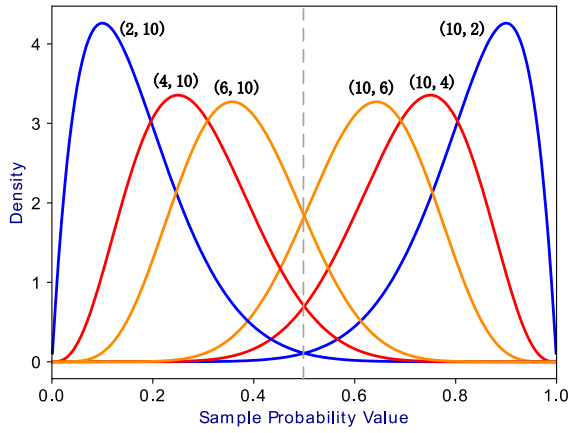


Fig. 6. Probability density curves of Beta distribution and the symmetric distribution under different (α, β) parameters. (α, β) parameters are marked near each curve.

As sampling distribution for grid map occupancy, Beta distribution has the following three advantages:

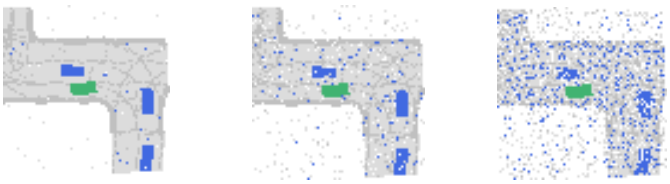
1) The value range of Beta distribution sampling is [0, 1], which meets the requirement of probability value;

2) Beta distribution is a conjugate prior distribution related to presence or absence, success or failure, such as Bernoulli distribution and Binomial distribution, which meets the requirements for describing BEV perception;

3) Beta distribution is very flexible. By changing the (α, β) parameters, it can simulate the Uniform distribution, Normal distribution, Bell-shaped distribution, U-shaped distribution, and many other distributions on [0, 1].

Therefore, in the experiment, we take Beta distribution as an example to sample SBEV-P data. When α is constant, the β mainly affects the mean of Beta distribution function. When β increases, the mean value will decrease, which corresponds to the fact when the actual occupancy of the grid is 1, the probability value assigned to the grid by the single vehicle perception is more deviated from 1. Similarly, when the actual occupancy of the grid is 0, we adopt the symmetric Beta distribution of the previous function, so that the probability value assigned by single vehicle perception will also be more deviated from 0. Taken together, it reduces the single vehicle perception ability. Fig. 6 shows the probability density curves of Beta distribution under different groups of symmetric parameter values. We set the α parameter to 10, and the β parameter to 2, 4, and 6 respectively to simulate the different perception ability.

As shown in Fig. 7, we compare the SBEV-I corresponding to the probability matrix under the three groups of parameters. It can be seen that with the increase of β parameter, the perception accuracy decreases, which is reflected in more scattered noises and more sparse objects in the corresponding BEV image. The parameter settings can simulate many cases related to vehicle perception, such as the equipment of low-precision Lidar, and the execution errors of sensors.



α=10, $\beta = 2$ α=10, $\beta = 4$ α=10, $\beta = 6$

Fig. 7. Different single vehicle perception ability of different Beta distribution parameters.

We set the initial value as $\alpha = 10, \beta = 4$. When the true occupancy state of the grid is 1, the mean and variance of Beta distribution are 0.71 and 0.01, representing CAVs with relatively strong perception ability. When the true occupancy state of the grid is 0, the distribution function used to sample the probability value is symmetric $Beta(4,10)$. The probability matrix can be used for generating initial SBEV-P and corresponding SBEV-I. The generated data from original datasets can be the simulation data of our experiments.

In our experiment, the CPU of the machine is Intel 10900X, and the GPU is RTX 3090. Our operating system is Ubuntu 18.04LTS with 128GB RAM.

For the training of the dataset, we set batchsize as 4, the maximum number of vehicles as $N=16$, and the embedding dimension as $E=64$. When the model is applied to other scenario datasets, the training parameters can be adjusted appropriately. The deep learning framework is Pytorch. During the training process, we select Adam optimizer with the initial learning rate of 0.001. The evaluation function and loss function are shown in *Section III.C.*, where the parameter is set as $\lambda_V=1, \lambda_M=0.03$. L1 regularization of $\lambda_R=1e-6$ coefficient is added to suppress overfitting. We train for about 150 epochs. The optimal model is selected according to the performance of the validation set, and the final model performance is tested on the test set. The training loss and validation set evaluation curves of the BEV-V2X model are shown in Fig. 8.

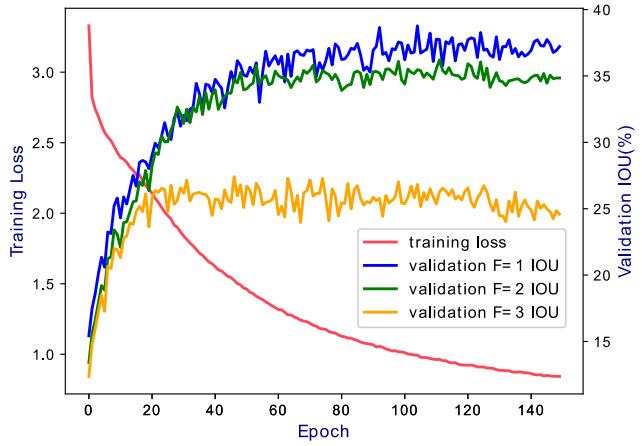


Fig. 8. Loss and validation IOU curve of training process.

B. *The Performance Testing Results of BEV-V2X*

We select typical neural network models MLP [72], LSTM [73], ResNet [61], and Casual Conv [64] to replace the attention core module in BEV-V2X. The rest of the embedding and output upsample layer remain unchanged, and then we compare them with the BEV-V2X model. In addition, we also ablate the attention module as BEV-V2X-S, BEV-V2X-T, BEV-V2X-S+T, and BEV-V2X-M+S+T respectively, which indicate the spatial attention, temporal attention, spatial and temporal attention, and spatial and temporal attention with global map. We then verify the effectiveness of each part of the attention module.

When evaluating the model effectiveness, we initially assume that all vehicles in the scenario are CAVs, i.e., the roadside unit can collect the local SBEV data of all vehicles. The IOU evaluation metric is used to compare the fusion and prediction performance of different types of models, and the average IOU values are shown in Table I.

Over time, we find that the grid occupancy IOU of the fusion and prediction models gradually decreases. The reason is that the historical data features have a limited time horizon for reference, and the error of the previous prediction will also accumulate. The final BEV-V2X-M+S+T model can reach the IOU value of 39, and the prediction can still reach the IOU value of 28 when the prediction time extends to 3s. Compared with the results of real-time scene segmentation and object detection related to BEV perception with high accuracy [12], [30], [48], the IOU value of BEV-V2X model is acceptable. Furthermore, considering that the fusion and prediction task of global CBEV is more difficult, the effectiveness of the proposed model is relatively good and valuable.

As shown in Table I, compared with non-attention models, when attention is used in the core module, it can better learn the spatiotemporal relationship of the data feature. In the future time stamps, the overall IOU value is better than that of non-attention models. For attention-based models, by comparing the ablation results of temporal, spatial and map modules, it can be found that the fusion and prediction result of spatiotemporal attention is better than that of temporal and spatial attention alone. If the prior map information is added to extract global spatial attention, the performance of the model can be further improved, which is also one of the advantages of the roadside deployment model for fusion and prediction. For the sake of clarity, the following BEV-V2X model refers to the overall BEV-V2X-M+S+T model.

As shown in Fig. 9 and Fig. 10, we select two typical examples from the test set to show the fusion and prediction results of the model. More examples can refer to the supplement video files. The model can well extract the motion and interaction information of each vehicle in the control area, and infer the future motion patterns and spatiotemporal changes. We take one vehicle as the ego at each timestamp and build the spatial attention weight of the other vehicles' corresponding perception area into the heat map. Then the he-

Table I
COMPARISON OF FUSION AND PREDICTION PERFORMANCE BY DIFFERENT METHODS
(F refers to future predicted timestamp, similarly hereinafter)

Type	*Method*	*IOU* (*F*=1s)	*IOU* (*F*=2s)	*IOU* (*F*=3s)
Comparison Models	*MLP*	33.9	31.4	24.7
	LSTM	30.5	30.4	27.0
	ResNet	30.8	30.1	25.4
	Casual Conv	34.9	32.9	24.0
Ablation Models	*BEV-V2X-S*	37.6	33.7	25.6
	BEV-V2X-T	36.0	34.3	26.7
	BEV-V2X-S+T	37.9	34.8	27.6
	BEV-V2X-M+S+T	**39.0**	**35.6**	**28.7**

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

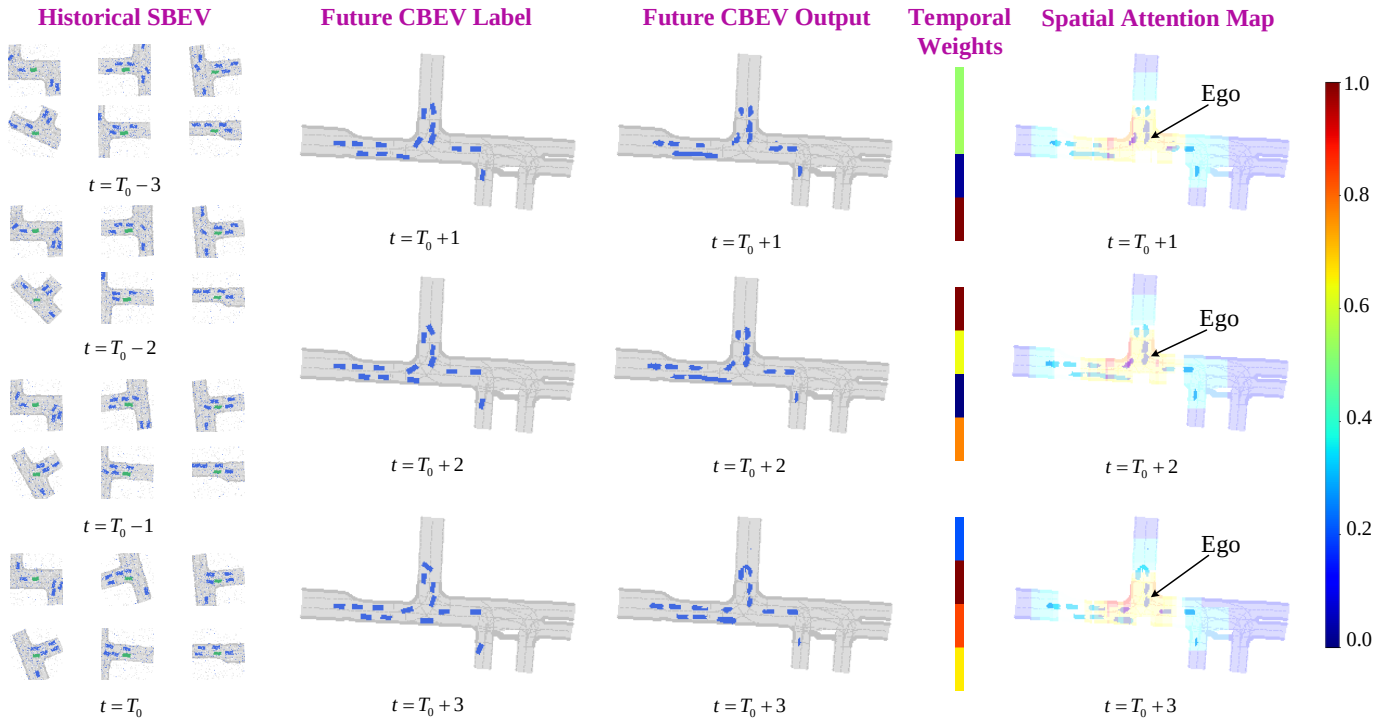


Fig. 9. Fusion and prediction example in dataset part 5, with attention weights visualization of ego vehicle id 8. (Due to the large number of vehicles in the area, only six typical historical SBEVs are listed here)

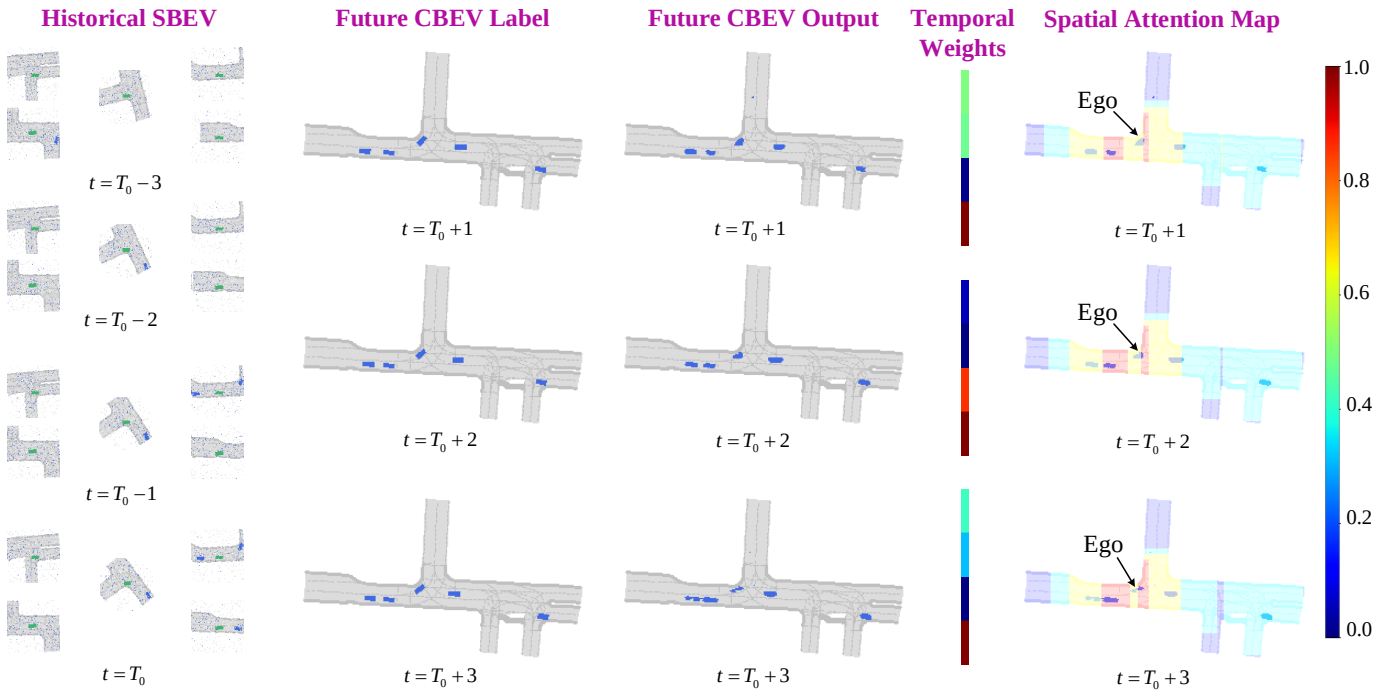


Fig. 10. Fusion and prediction example in dataset part 3, with attention weights visualization of ego vehicle id 81.

at map is overlaid on the global CBEV result to intuitively explain the spatial interaction that the model focuses on. In addition, we extract the temporal attention weights of the same ego vehicle on the historical data of the past 3s (4 sampling timestamps) and perform the visualization accordingly to show the temporal fusion.

For feature fusion at the temporal level, the attention-based model learns appropriate temporal weights corresponding to the historical 3s (4 sampling timestamps) data features via data-driven methods to achieve the temporal fusion.

As for the feature fusion at the spatial level, in Fig. 9, there are a large number of vehicles in the control area, and the

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

vehicle in the central area of the intersection is used to construct the spatial attention map as the ego vehicle. It can be seen that when the model fuses and predicts the relevant information of the ego, it mainly focuses on the intersection conflict area and the vehicles with interactions. In Fig. 10, the number of vehicles in the control area is relatively small, and the vehicle which is turning right is used to construct spatial attention map as ego vehicle. It can be seen that the model mainly focuses on the road areas and vehicles located on main road that have certain influence on its turns. The above focuses are consistent with human driving behaviors, indicating that the attention-based deep learning model can effectively learn relevant knowledge such as motion patterns and spatial interaction contained in naturalistic driving data.

C. *Comparison between Different Levels of CAV Rate*

When the cooperative driving scenario is not full of CAVs, we explore the ability of the proposed method to obtain the state of non-CAVs and the global BEV. We set the rate of CAVs in the scenario as 80%, 70%, and 60%, respectively. As shown in Fig. 11, we compare the IOU metric of CBEV results predicted by the BEV-V2X model.

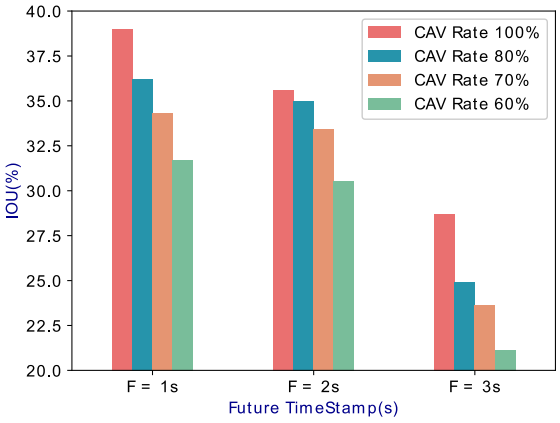


Fig. 11. Fusion and prediction IOU comparison of different CAV rates.

When the CAV rate is higher, the IOU of fusion and prediction is also relatively higher. The reason is that the local SBEV of CAVs collected by the roadside unit will have more overlapping areas. For grid state estimation, there are more data sources for reference, and the model with attention mechanism can be well applied to integrate.

With the decrease of CAV rate, the performance of fusion and prediction will decrease accordingly, and it will also decrease gradually with time. However, according to the results of previous scene segmentation works [12], [30], [48], the IOU metric of global CBEV output by the model is still maintained at a relatively high level in the future time horizon. It indicates that in the cooperative driving scenario where CAVs and normal vehicles are mixed, the BEV-V2X model based on V2X communication can still predict the spatiotemporal changes of the global scene very well by fusing the data transmitted by all CAVs.

As shown in Fig. 12, we show the fusion and prediction results when the CAV rate decreases. Thus, with the help of CAVs in the control area, the proposed model can still perceive other non-CAVs well, infer their future positions and states, and maintain good understanding and prediction ability.

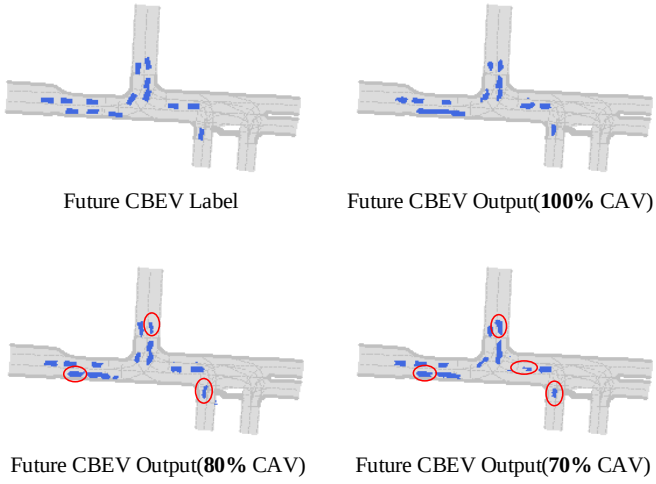


Fig. 12. Fusion and prediction example of different CAV rates. (Non-CAVs have been marked by red circles)

D. *Comparison between Perception Ability of Single Vehicles*

When the perception ability of each CAV is enhanced or weakened, we explore the impact of perception ability on the final CBEV fusion and prediction.

Table II
COMPARISON BETWEEN DIFFERENT PERCEPTION ABILITY OF SINGLE VEHICLES

IOU / *Beta Parameters*	*F*=1s	*F*=2s	*F*=3s
$\alpha=10,\beta=2$	**39.3**	**37.8**	**29.2**
$\alpha=10,\beta=4$	39.0	35.6	28.7
$\alpha=10,\beta=6$	35.3	32.5	26.3

As shown in *Section IV.A*, we conduct experiments for SBEV data with different Beta distribution parameters. The testing results are compared in Table II. It can be found that when the perception ability of single vehicle decreases, the IOU metric of fusion and prediction also decreases correspondingly. However, when the perception parameters are within certain ranges, the performance still maintains a relatively high level. The results suggest the following two points:

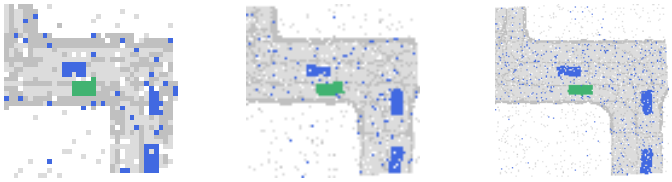
1) The fusion and prediction of multi-vehicle information collected via V2X communication can properly cope with the weaknesses of inaccurate and limited perception. Through multi-source data fusion, the accuracy of global information can be maintained at relatively high level even when the perception ability of vehicles is not high.

2) Cooperative driving fusion and prediction also depends on the single vehicle perception ability to a certain extent. If the perception ability of the vehicle is improved, V2X communication can play a more powerful role. However, once

the perception ability reaches a certain high level, for example, if the Beta parameter is changed from (10,4) to (10,2) to improve the perception, the final CBEV IOU metric will be slightly improved.

E. Comparison between Different Grid Size

We further explore the impact of different grid sizes on CBEV fusion and prediction. The fineness of the grid was set to 0.25m, 0.5m and 1.0m squares for each pixel respectively. We conduct experiments under the same Beta parameter to ensure the consistency of the perception ability. The corresponding SBEV-I differences can be visually compared in Fig. 13.



1.0*m* 0.5*m* 0.25*m*

Fig. 13. Different SBEV visualization of different grid size parameters.

For the model training with different grid sizes, the parameters shown in Fig. 3 are adapted accordingly. The comparisons of fusion and prediction performance are shown in Table III.

Table III
COMPARISON BETWEEN DIFFERENT GRID SIZE

IOU / *Grid Size*	*F*=1s	*F*=2s	*F*=3s
1.0*m*	35.2	34.2	25.4
0.5*m*	39.0	35.6	**28.7**
0.25*m*	**39.1**	**36.9**	28.1

As the grid size corresponding to BEV pixels decreases, i.e., the fineness increases, the IOU metric of fusion and prediction shows a trend of improvement. When the grid fineness is changed from 1.0m to 0.5m, the IOU increases significantly. While when the grid fineness is changed from 0.5m to 0.25m, the IOU increases slightly, and the prediction performance at different future timestamps cuts both ways. Furthermore, considering that as the fineness of the grid increases, the computation burden of the model will increase in quadratic order, so in the simulation environment of this paper, it is more appropriate to select 0.5m grid parameters for transmission, fusion and prediction. It is also compatible with the grid size settings in previous BEV-related researches [12], [13], [15].

V. Further Discussion about Methods via Single Vehicles, V2V Communication, and V2I Communication

There are mainly three different methods for fusion and prediction based on BEV data. One is that the single vehicle generates SBEV by its own perception system, and does not acquire other vehicles' information, so it only conducts prediction by the spatiotemporal features of its own SBEV data. Another is to transmit BEV among CAVs via V2V communication to fully utilize the shared information for spatiotemporal fusion and prediction. The third is the method adopted by our model, which applies roadside unit or cloud center to collect all CAVs' information in the control area in a unified and centralized manner for global fusion and prediction.

First, we compare the two methods based on V2V communication and V2I communication. Through the above experimental comparison, it can be seen that the application of edge computing based on V2I communication has the following obvious advantages:

1) The global map information of the control area can be pre-stored on the road side. As shown in Table I, when the map is input into the model as a priori, it is helpful for the attention module to better extract the global spatial position information of vehicles and improve the accuracy of model fusion and prediction. In addition, the map information can be directly used as the output of the occupancy state of road elements in the final CBEV result, which can improve the accuracy of environment perception.

2) Compared with vehicles, roadside units have a wider field of view and a wider range of information collection [74]. In addition, the application of data model [38] can efficiently store and extract CAVs' data. As shown in Fig. 11, as the number of CAVs increases, the model can integrate more information and achieve better results.

3) Compared with vehicles, roadside units have stronger computing resources [75] and can efficiently and quickly complete the deployment and calculation of models.

Further, to compare the fusion and prediction performance based on V2X approach and single vehicle, we conduct the experiment to predict the future global BEV (SBEV to CBEV) based on V2X communication and its own local BEV (SBEV to SBEV) based on its own local historical BEV data.

In the SBEV prediction experiment, we also set the perception parameter as $Beta(10,4)$. In the model shown in Fig. 3, the data input is the historical SBEV data, and the dimension is $(B,T,1,E)$. After the ResNet embedding operation, the features are input to the attention layer. Since the driving environment around the single vehicle will gradually change with the movements, the map-based global spatial aware attention is not set in the spatiotemporal feature extraction layer. The rest are the same as the modules of BEV-V2X. The experiment comparison results are shown in Table IV.

When only historical SBEV data are used to predict the future global CBEV, due to the limited perception range of the vehicle itself, the estimation and prediction of scenario elements outside its perception range will have large deviations, thus the IOU metric values will be very low. While the fusion and prediction based on V2X communication can collect more vehicles' information in a wider range. The spatial and temporal features of SBEV constructed by vehicles

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

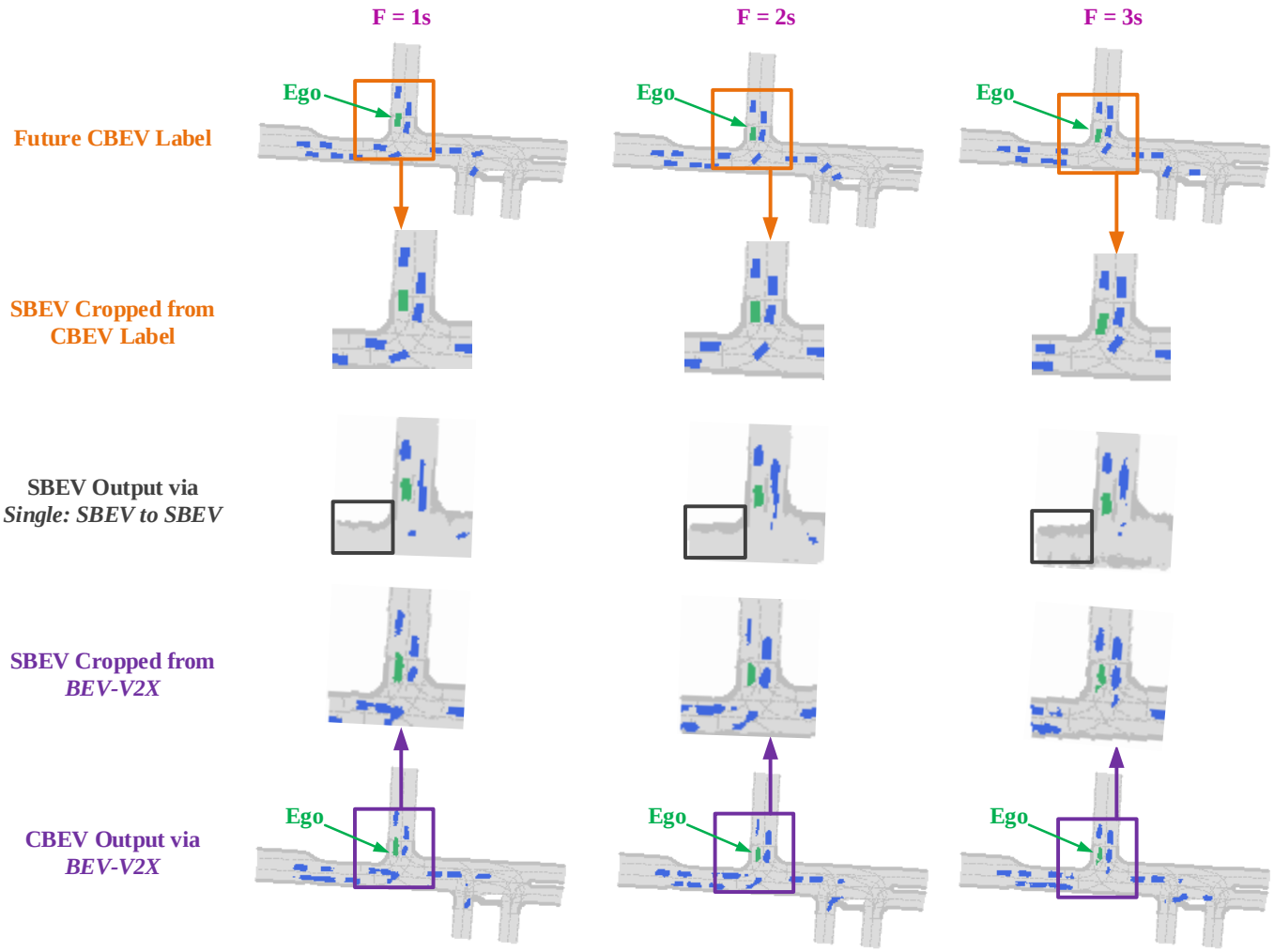


Fig. 14. Comparison example between SBEV output via *Single: SBEV to SBEV* model and corresponding SBEV cropped from *V2X: BEV-V2X* model.

are integrated by BEV-V2X model, which significantly expands the perception range of each vehicle, and the IOU values of global CBEV results also reach a high level.

When only historical SBEV data are used to predict future local SBEV information, we take SBEVs corresponding to the perception range of vehicles cropped from the global CBEV results based on V2X communication for comparison. Typical examples are shown in Fig. 14. Further combined with the results of Table IV, the following four points can be illustrated:

Table IV
COMPARISON BETWEEN METHODS VIA SINGLE VEHICLES AND V2X COMMUNICATION

Methods \ *IOU*	*F*=1s	*F*=2s	*F*=3s
***Single**: SBEV to CBEV*	19.0	18.1	16.8
***V2X**: BEV-V2X*	**39.0**	**35.6**	**28.7**
***Single**: SBEV to SBEV*	36.0	22.6	17.0
***V2X**: BEV-V2X crop SBEV*	**41.4**	**39.3**	**32.3**

1) Since the data source of *Single: SBEV to SBEV* is only the vehicle itself, there are more uncertainties in grid occupancy states. While the V2X-based approach can collect the historical data of many vehicles, so the methods can expand the perception area significantly. In addition, there are also overlapping parts among the corresponding perception areas of the single vehicles, which can be effectively applied to the fusion. The results in Table IV show that the IOU metric value of fusion and prediction is significantly improved. Furthermore, from Fig. 14, it can be seen that the fusion and prediction performance of the cropped local BEV of the BEV-V2X model is significantly better than that of the single vehicle-based prediction.

2) The prediction via single vehicle cannot capture the information of vehicles that will enter the perception area of single vehicle in the future. As shown in Fig. 14, in the results from *Single: SBEV to SBEV*, the grid occupancy in the lower left black box area are not correctly predicted. The true situation is that although no vehicle occupied the lower left part of the ego SBEV area in the historical time horizon (the ego vehicle was on the branch lane, and the perception was difficult to touch the main lane), there will exist vehicles

entering the perception area of the ego vehicle in the future 3s. Therefore, it is difficult for vehicles to capture relevant information only from the historical data of single vehicle. Fortunately, with the help of V2X communication, it can perceive the global vehicles' movements information to make better fusion and prediction.

3) When we deploy BEV-V2X model via V2X communication on the roadside/cloud, the global map information of the control area can be pre-stored and directly applied to output results. In contrast, the fusion and prediction method based on single vehicle not only predicts the occupancy state of vehicles in the grid map, but also needs to predict the occupancy of road elements such as drivable areas and road markings. As shown in Fig. 14, the results show certain deviations compared with the real road map, which weakens the accuracy of environment perception.

4) As shown in Table IV, the fusion and prediction index based on single vehicle decreases rapidly with time, while the index of methods based on V2X communication decreases slowly. It shows obvious advantages in long-term prediction tasks.

VI. CONCLUSIONS

In this paper, a BEV fusion and prediction method based on V2X communication and attention neural network model is proposed. The experiment results show that the performance of the proposed BEV-V2X model is significantly better than the prediction based on single vehicle perception. When the grid size is 0.5m, the fusion and prediction accuracy of the global CBEV at the future 1s timestamp can reach 40 IOU index, and the accuracy of the future 3s timestamp can still reach 30 IOU index. Referring to the high accuracy results of BEV perception related real-time scene segmentation studies, the BEV-V2X model can meet the requirements of accurate scene understanding and inference. Even if the cooperative driving scenario is not full of CAV, when the CAV rate reaches over 60%, the fusion and prediction method based on V2X communication can maintain the IOU index of more than 30, and the model can still accurately estimate and predict the spatiotemporal dynamic changes of the global scenario. It will also be better with the enhancement of single vehicle perception ability. The fusion and prediction results can further support driving safety warning, multi-vehicle cooperative driving, cooperative traffic control, and other applications.

Due to the space limitation of this paper, the following three aspects are not further discussed:

First, in the cooperative driving scenario, the roadside unit collects CAV data in the control area to obtain global information. BEV data contain more semantic information, and can be applied to the driving scenario where CAVs and non-CAVs are mixed. However, BEV data also has some disadvantages compared to trajectory data, such as complex data structure and large space consumption, which may lead to slow transmission and computation. In the future, we will further analyze the advantages and disadvantages, and the influence of interactive data format between vehicle side and road side in the researches of cooperative warning and planning.

Second, through the above experiments, it can be observed that when the CAV rate increases or the single vehicle perception ability increases, the fusion and prediction performance will be improved. However, when the single vehicle perception ability reaches a certain high level, the performance is not significantly improved, while increasing the CAV rate will achieve more significant improvement. In addition, there also exist problems such as visual blind area and vehicle occlusion in vehicle perception. The limited ability will increase the bottleneck of perception ability improvement. The research in this paper will prompt us to further think about measures to promote the development of autonomous driving, focusing on increasing the CAV rate or improving the perception ability, which will also be an interesting topic for researchers to discuss.

Third, in this paper, we focus on demonstrating the experiments of typical intersection scenario datasets. According to studies on motion prediction [15]-[17] and time statistics of datasets [68], we set the historical time and future time horizon as appropriate values. Under different types of driving scenarios and behaviors, the corresponding time parameter settings should also be changed and adapted. We will further discuss the issue in the future work related to driving behaviors simulation and verification.

REFERENCES

[1] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[2] J. Liu *et al*., "Deep Instance Segmentation with Automotive Radar Detection Points," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 84-94, 2023.

[3] M. Scholtes *et al*., "6-Layer Model for a Structured Description and Categorization of Urban Traffic and Environment," *IEEE Access*, vol. 9, pp. 59131-59147, 2021.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision Meets Robotics: The Kitti Dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[5] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes Dataset for Semantic Urban Scene Understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp.3213–3223.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu,A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "Nuscenes: A Multimodal Dataset for Autonomous Driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11621–11631.

[7] L. Fang, D. Zhu, J. Yue, B. Zhang, and M. He, "Geometric-Spectral Reconstruction Learning for Multi-Source Open-Set Classification with Hyperspectral and Lidar Data," in *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 10, pp. 1892-1895, 2022.

[8] Y. Sun, J. Li, Y. Wang, X. Xu, X. Yang, and Z. Sun, "ATOP: An Attention-to-Optimization Approach for Automatic LiDAR-Camera Calibration via Cross-Modal Object Matching," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 696-708, 2023.

[9] O. Natan and J. Miura, "End-to-End Autonomous Driving with Semantic Depth Cloud Mapping and Multi-Agent," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 557-571, 2023.

[10] Z. Wu *et al*., "Surround-View Fisheye BEV-Perception for Valet Parking: Dataset, Baseline and Distortion-Insensitive Multi-Task Framework," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2037-2048, 2023.

[11] S. Fadadu, S. Pandey, D. Hegde *et al*., "Multi-view Fusion of Sensor Data for Improved Perception and Prediction in Autonomous Driving,"

in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2349-2357.

[12] T. Roddick and R. Cipolla, "Predicting Semantic Map Representations from Images using Pyramid Occupancy Networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, p. 11.

[13] Z. Li, W. Wang, H. Li *et al.*, "BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers," *arXiv preprint arXiv:2203.17270*, 2022.

[14] C. Chang *et al*., "MetaScenario: A Framework for Driving Scenario Data Description, Storage and Indexing," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1156-1175, 2023.

[15] K. Zhang, L. Zhao, C. Dong, L. Wu, and L. Zheng, "AI-TP: Attention-Based Interaction-Aware Trajectory Prediction for Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 73-83, 2023.

[16] J. Zhao, T. Qu, X. Gong, and H. Chen, "Interaction-Aware Personalized Trajectory Prediction for Traffic Participant Based on Interactive Multiple Model," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2184-2196, 2023.

[17] L. Fang, Q. Jiang, J. Shi *et al.*, "TPNet: Trajectory Proposal Network for Motion Prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 6797-6806.

[18] H. Pei, J. Zhang, Y. Zhang, X. Pei, S. Feng, and L. Li, "Fault-Tolerant Cooperative Driving at Signal-Free Intersections," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 121-134, 2023.

[19] J. Zhang, H. Pei, X J Ban *et al*., "Analysis of Cooperative Driving Strategies at Road Network Level with Macroscopic Fundamental Diagram," *Transportation Research Part C: Emerging Technologies*, vol. 135, pp. 103503, 2022.

[20] Y. Guo, D. Yao, B. Li, H. Gao, and L. Li, "Down-Sized Initialization for Optimization-Based Unstructured Trajectory Planning by Only Optimizing Critical Variables," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 709-720, 2023.

[21] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A Unified Model to Map, Perceive, Predict and Plan", in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 14403-14412.

[22] L. Li, W. -L. Huang, Y. Liu, N. -N. Zheng, and F. -Y. Wang, "Intelligence Testing for Autonomous Vehicles: A New Approach," *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 2, pp. 158-166, 2016.

[23] L. Li *et al*., "Parallel Testing of Vehicle Intelligence via Virtual-Real Interaction," *Sci. Robot.*, vol. 4, no. 28, 2019.

[24] X. Li, S. Teng, B. Liu, X. Dai, X. Na, and F. -Y. Wang, "Advanced Scenario Generation for Calibration and Verification of Autonomous Vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 5, pp. 3211-3216, 2023.

[25] H. Pei, Y. Zhang, Q. Tao, S. Feng, and L. Li, "Distributed Cooperative Driving in Multi-Intersection Road Networks," *IEEE Transactions on Vehicular Technology*, vol. 70., no. 6, pp. 5390-5403, 2021.

[26] Z. Wang, K. Han, and P. Tiwari, "Digital Twin-Assisted Cooperative Driving at Non-Signalized Intersections," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 2, pp. 198-209, 2022.

[27] S. Chen, J. Hu, Y. Shi, L. Zhao, and W. Li, "A Vision of C-V2X: Technologies, Field Testing, and Challenges with Chinese Development," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3872–3881, 2020.

[28] A. M. Muad, A. Hussain, S. A. Samad, M. M. Mustaffa, and B. Y. Majlis, "Implementation of Inverse Perspective Mapping Algorithm for the Development of an Automatic Lane Tracking System," in *Proc. IEEE TENCON Region 10 Conf.*, 2004, pp. 207–210.

[29] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Crossview Semantic Segmentation for Sensing Surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

[30] A. Saha, O. Mendez, C. Russell *et al*., "Translating Images into Maps," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9200-9206.

[31] S. Mandal, S. Biswas, V. E. Balas, R. N. Shaw, and A. Ghosh, "Motion Prediction for Autonomous Vehicles from Lyft Dataset using Deep Learning," in *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, 2020, pp. 768-773.

[32] S. Narayanan, R. Moslemi, F. Pittaluga, B. Liu, and M. Chandraker, "Divide-and-Conquer for Lane-Aware Diverse Trajectory Prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15799-15808.

[33] K. Zhang, C. Chang, W. Zhong, S. Li, Z. Li, and L. Li, "A Systematic Solution of Human Driving Behavior Modeling and Simulation for Automated Vehicle Studies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 21944-21958, 2022.

[34] J. Kim, R. Mahjourian, S. Ettinger, M. Bansal, B. White, B. Sapp, and D. Anguelov, "StopNet: Scalable Trajectory and Occupancy Prediction for Urban Autonomous Driving," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 8957-8963.

[35] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A Survey on Trajectory-Prediction Methods for Autonomous Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652-674, 2022.

[36] A. Laddha, S. Gautam, G. P. Meyer, C. Vallespi-Gonzalez, and C. K. Wellington, "RV-FuseNet: Range View Based Fusion of Time-Series LiDAR Data for Joint 3D Object Detection and Motion Forecasting,"in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, 2021, pp. 7060-7066.

[37] S. Tsugawa, "Inter-Vehicle Communications and their Applications to Intelligent Vehicles: An Overview," in *Proceedings of IEEE Intelligent Vehicles Symposium*, vol. 2, pp. 564-569, 2002.

[38] H. Yu, C. Chang, S. Li, and L. Li, "CD-DB: A Data Storage Model for Cooperative Driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 492-501, 2023.

[39] K. Tischler and B. Hummel, "Enhanced Environmental Perception by Inter-Vehicle Data Exchange," in *Proc. IEEE Intell. Veh. Symp.*, 2005, pp. 313–318.

[40] S. Fujii, A. Fujita, T. Umedu, H. Yamaguchi, T. Higashino, S. Kaneda, and M. Takai, "Cooperative Vehicle Positioning via V2V Communications and Onboard Sensors," in *Proc. IEEE 74th VTC*, Sep. 2011, pp. 1–5.

[41] S. J. Greybush, E. Kalnay, T. Miyoshi, K. Ide, and B. R. Hunt, "Balance and Ensemble Kalman Filter Localization Techniques," *Monthly Weather Review*, vol, 139, no. 2, pp. 511-522, 2011.

[42] H. Tan and J. Huang, "DGPS-Based Vehicle-to-Vehicle Cooperative Collision Warning: Engineering Feasibility Viewpoints," *IEEE Transactions on Intelligent Transportation Systems*, vol. 7, no. 4, pp. 415-428, 2006.

[43] A. Shetty, M. Yu, A. Kurzhanskiy, O. Grembek, H. Tavafoghi, and P. Varaiya, "Safety Challenges for Autonomous Vehicles in the Absence of Connectivity," *Transportation Research Part C: Emerging Technologies*, vol.128, pp. 103133, 2021.

[44] M. Ramezani, J. A. Machado, A. Skabardonis, and N. Geroliminis, "Capacity and Delay Analysis of Arterials with Mixed Autonomous and Human-Driven Vehicles," in *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Naples, Italy, 2017, pp. 280-284.

[45] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, Dallas, TX, USA, 2019, pp. 514-524.

[46] H. Qiu, P. Huang, N. Asavisanu, X. Liu, K. Psounis, and R. Govindan, "Autocast: Scalable Infrastructure-Less Cooperative Perception for Distributed Collaborative Driving," *arXiv preprint arXiv:2112.14947*, 2021.

[47] R. Mao, J. Guo, Y. Jia, Y. Sun, S. Zhou, and Z. Niu, "DOLPHINS: Dataset for Collaborative Perception enabled Harmonious and Interconnected Self-driving," in *Proceedings of the Asian Conference on Computer Vision*, pp. 4361-4377, 2022.

[48] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, 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)*, 2022, pp. 21361-21370.

[49] S. Ren, S. Chen, and W. Zhang, "Collaborative Perception for Autonomous Driving: Current Status and Future Trend," in *Proceedings of 2021 5th Chinese Conference on Swarm Intelligence and Cooperative Control*, Springer, Singapore, 2023, pp. 682-692.

[50] 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 *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2021, pp. 88-100.

[51] T. H. Wang, S. Manivasagam, M. Liang *et al*., "V2VNet: Vehicle-to-Vehicle Communication for Joint Perception and Prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, Cham, 2020, pp. 605-621.

[52] R. Xu, Z. Tu, H. Xiang *et al*., "CoBEVT: Cooperative Bird's Eye View Semantic Segmentation with Sparse Transformers," *arXiv preprint arXiv:2207.02202*, 2022.

[53] R. Xu, H. Xiang, Z. Tu *et al*., "V2X-ViT: Vehicle-to-Everything Cooperative Perception with Vision Transformer," *arXiv preprint arXiv:2203.10638*, 2022.

[54] 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 *2022 International Conference on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, 2022, pp. 2583-2589.

[55] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning Distilled Collaboration Graph for Multi-Agent Perception," in *Advances in Neural Information Processing Systems*, 2021, pp. 29541-29552.

[56] M. Fadda, M. Murroni, and V. Popescu, "Interference Issues for VANET Communications in the TVWS in Urban Environments," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 4952-4958, 2016.

[57] X. Li, P. Ye, J. Li, Z. Liu, L. Cao, and F. -Y. Wang, "From Features Engineering to Scenarios Engineering for Trustworthy AI: I&I, C&C, and V&V," *IEEE Intelligent Systems*, vol. 37, no. 4, pp. 18-26, 2022.

[58] C. Chang, K. Zhang, J. Zhang, S. Li, and L. Li, "Driving Safety Monitoring and Warning for Connected and Automated Vehicles via Edge Computing," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 2022.

[59] A. Vaswani *et al.*, "Attention is All You Need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[60] K. Chen, G. Chen, D. Xu, L. Zhang, Y. Huang, and A. Knoll. "NAST: Non-Autoregressive Spatial-Temporal Transformer for Time Series Forecasting," *arXiv preprint arXiv:2102.05624*, 2021.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385*, 2015.

[62] A. Oord, S. Dieleman, H. Zen *et al.*, "Wavenet: A Generative Model for Raw Audio," *arXiv preprint arXiv:1609.03499*, 2016.

[63] F. Yu and V. Koltun, "Multi-scale Context Aggregation by Dilated Convolution," *arXiv preprint arXiv:1511.07122*, 2015.

[64] Y. Abu Farha, and J. Gall, "Ms-tcn: Multi-Stage Temporal Convolutional Network for Action Segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[65] US Department of Transportation, "NGSIM – Next Generation Simulation,", 2007. [Online]. Available: http://www.ngsim.fhwa.dot.gov.

[66] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The Highd Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems,"in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2118–2125.

[67] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hart-nett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse:3D Tracking and Forecasting with Rich Maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8748–8757.

[68] W. Zhan *et al.*, "Interaction Dataset: An International, Adversarial and Cooperative Motion Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv preprint arXiv:1910.03088*, 2019.

[69] K. Rakesh, "VANET parameters and applications: a review," *Global Journal of Computer Science and Technology*, vol. 10, no. 7, 2010.

[70] G. D. Tipaldi, and K. O. Arras, "FLIRT - Interest Regions for 2D Range Data," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 3616-3622.

[71] J. Clemens, T. Kluth, and T. Reineking, "β-SLAM: Simultaneous Localization and Grid Mapping with Beta Distributions," *Information Fusion*, vol. 52, 2019, pp. 62-75.

[72] H. Taud and J. F. Mas, "Multilayer Perceptron (MLP)," *Geomatic Approaches for Modeling Land Change Scenarios*, Springer, Cham, 2018, pp. 451-455.

[73] Y. Yu, X. Si, C. Hu, and J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, 2019.

[74] L. Jiang, T. G. Molnár, and G. Orosz, "On the Deployment of V2X Roadside Units for Traffic Prediction," *Transportation Research Part C: Emerging Technologies*, vol. 129, pp. 103238, 2021.

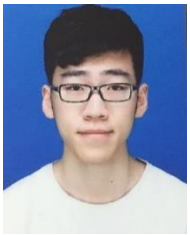
[75] C. Zhang, X. Lin, R. Lu, and P. -. Ho, "RAISE: An Efficient RSU-Aided Message Authentication Scheme in Vehicular Communication Networks," in *2008 IEEE International Conference on Communications*, 2008, pp. 1451-1457.

Cheng Chang received the B.S. degree from Tsinghua University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the Department of Automation. His current research interests include intelligent transportation systems, intelligent vehicles and machine learning.

Jiawei Zhang received the B.S. degree from Tsinghua University, Beijing, China, in 2020. He is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include autonomous driving, intelligent transportation systems, and deep reinforcement learning.

Kunpeng Zhang received the Ph.D. degree from the College of Mechanical and Vehicle Engineering, Hunan University, China, in 2019. From 2016 to 2017, he spent one year as a Joint Doctoral Student with the University of Michigan, Ann Arbor, MI, USA. He is currently a Post-Doctoral Researcher with the Department of Automation, Tsinghua University, Beijing, and a Lecturer with the College of Electrical Engineering, Henan University of Technology, Zhengzhou, China. His research interests include intelligent transportation systems and autonomous driving.

Wenqin Zhong received the B.S. degree in vehicle engineering from the School of Vehicle and Mobility, Tsinghua University, Beijing, China, in 2021. She is currently pursuing the master's degree with the Tsinghua Shenzhen International Graduate School. Her research interests include mixed traffic flow scheduling, and multi-vehicle cooperative controlling at intersections.



Xinyu Peng is currently pursuing the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include machine learning, intelligent transportation systems, and intelligent vehicle.

This article has been accepted for publication in IEEE Transactions on Intelligent Vehicles. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TIV.2023.3293954

Shen Li received the Ph.D. degree at the University of Wisconsin – Madison in 2018. He is a research associate at Tsinghua University. His research is about Intelligent Transportation Systems (ITS), Architecture Design of CAVH System, Vehicle-infrastructure Cooperative Planning and Decision Method, Traffic Data Mining based on Cellular Data, and Traffic Operations and Management.

Li Li (Fellow, IEEE) is currently a Professor with the Department of Automation, Tsinghua University, Beijing, China, where he was involved in artificial intelligence, intelligent control and sensing, intelligent transportation systems, and intelligent vehicles. He has authored over 140 SCI-indexed international journal articles and over 70 international conference papers. He was a member of the Editorial Advisory Board for *Transportation Research Part C: Emerging Technologies*, and a member of the Editorial Board of *Transport Reviews* and *Acta Automatica Sinica*. He also serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and IEEE TRANSACTIONS ON INTELLIGENT VEHICLES.