Multimodal-XAD: Explainable Autonomous Driving Based on Multimodal Environment Descriptions

Yuchao Feng, *Student Member, IEEE*, Zhen Feng, *Member, IEEE*, Wei Hua, and Yuxiang Sun, *Member, IEEE*

***Abstract*— In recent years, deep learning-based end-to-end autonomous driving has become increasingly popular. However, deep neural networks are like black boxes. Their outputs are generally not explainable, making them not reliable to be used in real-world environments. To provide a solution to this problem, we propose an explainable deep neural network that jointly predicts driving actions and multimodal environment descriptions of traffic scenes, including bird-eye-view (BEV) maps and natural-language environment descriptions. In this network, both the context information from BEV perception and the local information from semantic perception are considered before producing the driving actions and natural-language environment descriptions. To evaluate our network, we build a new dataset with hand-labelled ground truth for driving actions and multimodal environment descriptions. Experimental results show that the combination of context information and local information enhances the prediction performance of driving action and environment description, thereby improving the safety and explainability of our end-to-end autonomous driving network.**

***Index Terms*— Autonomous driving, decision making, multimodal explanations, BEV perception, explainable AI (XAI).**

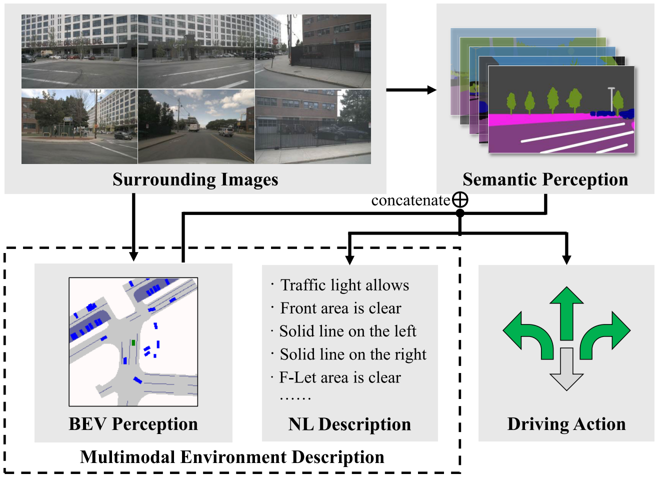


Fig. 1. The overview of our proposed method. The NL Description refers to the natural-language environment descriptions.

I. Introduction

OVER the past years, the research on autonomous driving has made great progress due to the impressive advancement of deep learning technologies [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. However, most existing deep learning-based autonomous driving networks suffer from the issue of lacking explainability, because deep neural networks are like black boxes. Without explainable control commands, it is unsafe to deploy these technologies in real-world environments. To provide explanations to autonomous driving networks, many methods [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] have been proposed. The explanations provided by these methods can be generally categorized into two types: visual explanation [12], [13], [14], [15], [16], [17] and natural-language explanation [18], [19], [20], [21], [22], [23], [24], [25], [26]. The visual explanation usually explains the network outputs by visualizing the inner process of the network [27], [28], [29], such as saliency maps and attention heat maps. The natural-language explanation [18], [19], [20], [21], [22], [23], [24], [25], [26] explains the network outputs in the form of phrases, such as driving action reasons and goals. Compared with the visual explanation, the natural-language explanation is easier to interpret and can give end users a better understanding of what triggers a particular behavior [18]. However, the natural-language explanation lacks the ability to describe how the inner process of the network works. Therefore, combining both visual and natural-language explanations may be a more effective way to explain the outputs of autonomous driving networks. So, in this work, we propose an explainable deep neural network that jointly predicts driving actions and the corresponding explanations in multimodal formats, including bird-eye-view (BEV) maps and natural-language environment descriptions for traffic scenes.

Recently, BEV perception for traffic scenes has attracted great attention because the BEV map is very straightforward for many downstream tasks [30], such as motion planning, behavioral intention prediction [31], etc. The existing methods on BEV perception can be generally divided into three

Received 24 February 2023; revised 19 January 2024, 9 May 2024, and 18 July 2024; accepted 11 September 2024. Date of publication 7 October 2024; date of current version 27 November 2024. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by Zhejiang Laboratory under grant 2021NL0AB01, and in part by the City University of Hong Kong under Grant 9610675. The Associate Editor for this article was B. Fidan. *(Corresponding author: Yuxiang Sun.)*

Yuchao Feng and Zhen Feng are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: yuchao.feng@connect.polyu.hk; zhen.feng@connect.polyu.hk).

Wei Hua is with Zhejiang Laboratory, Hangzhou, Zhejiang 311121, China (e-mail: huawei@zhejianglab.com).

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong (e-mail: yx.sun@cityu.edu.hk).

Digital Object Identifier 10.1109/TITS.2024.3467175

1558-0016 © 2024 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

categories according to the used sensors: point cloud-based methods [32], [33], [34], [35], vision-based methods [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48] and multimodal methods [49], [50], [51], [52]. The point cloud-based methods use point clouds produced by LiDAR or radar. The vision-based methods use RGB images produced by visual cameras. They need to transform the visual information from the Perspective View (PV) to BEV. The multimodal methods integrate data from various sensors, such as cameras, LiDAR and radar, to obtain BEV perception. Compared with the point cloud-based and multimodal methods, the vision-based methods are cost-effective but with relatively inferior semantic perception performance [30]. Moreover, the unsatisfactory BEV segmentation results of vision-based methods may further influence the downstream tasks. In other words, error accumulations may happen to the downstream tasks of BEV perception.

To alleviate the error accumulation, our proposed method considers both the context information from BEV perception and the local information from semantic perception (i.e., the semantic segmentation from surrounding images) before predicting driving actions and natural-language environment descriptions. Fig. 1 shows the overview of our proposed method. To train and evaluate our network, we release a new dataset containing $12,000$ image sequences. Each sequence includes images from surrounding cameras and the hand-labelled ground truth for driving actions, as well as multimodal descriptions of traffic scenes. The experimental results show that the combination of context information and local information improves the prediction performance of both driving actions and environment descriptions, which increases the safety and explainability of the autonomous driving network. The contributions of this work are summarized as follows:

1) A novel explainable decision-making network for autonomous driving is proposed by considering both the context information from BEV perception and local information from semantic perception.
2) The multimodal environment descriptions of traffic scenes, including BEV maps and natural-language environment descriptions, are applied to explain the driving actions.
3) A large-scale dataset containing $12,000$ image sequences is released. Each image sequence contains the hand-labelled ground truth for driving actions, as well as BEV maps and natural-language environment descriptions of traffic scenes.
4) The superiority of our proposed network over other networks is demonstrated on both our released dataset and the publicly available dataset. Our code and dataset are publicly available.[1]

The remainder of this paper is structured as follows. Section II reviews the related work. Section III presents the details of our proposed network. Section IV discusses the experimental results. Conclusions and future work are drawn in the last section.

II. Related Work

A. Explainable Autonomous Driving Networks

As aforementioned, two main streams of methods, including the visual explanation [12], [13], [14], [15], [16], [17] and natural-language explanation [18], [19], [20], [21], [22], [23], [24], [25], [26], are applied in explainable autonomous driving. The visual explanation is obtained by visualizing the inner process of the network. For example, Kim et al. [13] proposed to use the visual attention heat map to highlight regions that causally influence driving actions. Renz et al. [14] proposed an explainable planning transformer, called PlanT. In this network, by extracting and visualizing the attention weights, objects that are relevant and crucial for the agent's decision could be identified to increase the explainability. The limitation of visual explanations is that they are not easy to understand, especially for end users.

Besides the visual explanation, natural-language explanation has also been applied in some explainable autonomous driving networks [18], [19], [20], [21], [22], [23], [24], [25], [26]. For example, Xu et al. [19] proposed a deep learning-based network to jointly predict driving actions along with the corresponding natural-language reasons. While natural-language explanations serve as effective aids in understanding the decisions made by networks, they often fall short of revealing the inner process of networks. Taking into account the limitations of the visual and natural-language explanations, we believe that the multimodal explanation can better enhance the explainability of networks.

B. Vision-Based BEV Perception

In recent years, many efforts [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48] have been made in the field of vision-based BEV perception. For example, Kim and Kum [36] proposed a method that utilizes inverse perspective mapping to estimate the distance from a single monocular image by assuming that all image pixels are on the ground. However, the hard assumption that all pixels are on the ground sacrifices the height discrimination. To address this issue, Philion and Fidler [39] proposed a network to infer BEV representations from arbitrary camera rigs. Pan et al. [42] proposed the View Parsing Network (VPN) to parse the first-view observations into a BEV semantic map. Zhou et al. [46] proposed the Cross-View Transformers (CVT), an efficient attention-based model for map-view semantic segmentation from multiple cameras. Recently, to increase the robustness of vision-based BEV perception, Chen et al. [48] proposed the Masked BEV (M-BEV) perception framework to address the emergence of camera crashes. Experimental results show that the M-BEV framework significantly increases the performance of the different models for various missing camera emergencies.

C. Multimodality in Autonomous Driving

Multimodality has received great attention in autonomous driving, because it could enhance the perception and decision-making capabilities of autonomous vehicles [49],

[1] https://github.com/lab-sun/Multimodal-XAD

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

[50], [51], [52], [53], [54], [55], [56], [57], [58]. For example, Prakash et al. [57] proposed a novel multimodal fusion transformer, utilizing attention to integrate image and LiDAR representations. Fang et al. [58] proposed a cognitive accident prediction method by explicitly considering both the driver attention and text descriptions of traffic scenes. Despite these efforts in utilizing multimodality in autonomous driving, multimodality in existing works is in terms of inputs, there is still a research gap on how to utilize multimodality to enhance the explainability of end-to-end driving networks, which deserves further investigation. So, in this work, we propose to use multimodal environment descriptions as outputs to explain the driving actions.

D. Language-Driven Autonomous Driving Systems

Recently, with the rapid development of Large Language Models (LLMs), there has been a notable trend towards integrating LLMs into autonomous driving systems [59], [60], [61], [62], [63]. It is believed that the generalization ability of autonomous driving systems can be improved by utilizing the commonsense reasoning of the LLMs [63]. Moreover, the capability of LLMs in natural-language understanding and generation can be used to enhance the explainability of autonomous driving by producing contextually rich explanations [61]. For example, Sima et al. [62] investigated how to integrate the vision-language models (VLMs) into end-to-end driving systems to improve the generalization. Fu et al. [63] proposed to use the LLMs to understand traffic scenes in a human-like manner and assess their capabilities to reason, interpret, and memorize in complex scenarios.

E. Datasets for Autonomous Driving

So far, many real-world autonomous driving datasets have been released, such as KITTI [64], CityScapes [65], Apolloscape [66], nuScenes [67], BDD100K [68], etc. However, these datasets cannot encompass all situations, especially for some corner cases and long-tail scenarios. To generate specific scenarios at low cost, some simulators [69], [70] and world models [71], [72] are proposed, by which the synthetic datasets can be generated. For example, DriveDreamer [71] is able to generate high-quality driving videos of realistic traffic scenes and formulate reasonable driving policies. However, whether it is real-world datasets or synthetic datasets, they cannot be used for training explainable autonomous driving networks. To develop and verify explainable networks, some explainable datasets [18], [19], [26], [73], [74] have been proposed. The limitations of these explainable datasets are that they only contain natural-language explanations. To further increase the explainability, we build a new dataset that contains driving actions and multimodal environment descriptions. The details of the proposed dataset are provided in section IV-A.

III. The Proposed Network

A. The Network Architecture

As shown in Fig. 2, our proposed network, named Multimodal-XAD, mainly consists of five components: encoder, BEV module, semantic understanding (S-U) module, context embedding (C-E) module, and action-description (A-D) module. The network takes as input the images along with the intrinsics and extrinsics from surrounding monocular cameras (including the front camera, front left camera, front right camera, back camera, back left camera, and back right camera) to predict driving actions along with the multimodal environment descriptions, namely, BEV maps and natural-language environment descriptions of traffic scenes.

Let $X_i[k]$ denote the k-th image in the sequence $X_i[1:n]$ with an intrinsic matrix $I_i[k] \in \mathbb{R}^{3\times 3}$ and an extrinsic matrix $E_i[k] \in \mathbb{R}^{3\times 4}$, where X_i is the image sequence with several images from different surrounding cameras, i is the index of the image sequence in the dataset, n is the number of surrounding cameras. Here, we have $X_i[k] \in \mathbb{R}^{h\times w\times c}$, where h, w and c denote height, width and number of channels for an image, respectively. Given that many traffic scenes are complex, multiple driving actions may be applicable. So, Multimodal-XAD is designed to predict multiple driving actions A_i. Here, 4 categories of driving actions are adopted, including "move forward", "turn left/change to left lane", "turn right/change to right lane", and "stop/slow down". So, the driving action can be denoted as $A_i \in \{0, 1\}^4$.

For the multimodal environment descriptions of traffic scenes, both BEV maps D_i^{bev} and the natural-language environment descriptions D_i^{nl} are predicted. The BEV map is defined as the multi-class semantic BEV grid map of traffic scenes. The size and resolution of the BEV map are 100 meter × 100 meter and 0.5 meter × 0.5 meter, respectively. The number of semantic classes is 4, including road, vehicle, road/lane divider and background. So, BEV maps can be denoted as $D_i^{\text{bev}} \in \{0, 1, 2, 3\}^{200\times 200}$. The natural-language environment description D_i^{nl} contains 8 categories, including "traffic light allows", "front area is clear", "solid line on the left", "solid line on the right", "front left area is clear", "back left area is clear", "front right area is clear", and "back right area is clear". So, the natural-language environment description can be denoted as $D_i^{\text{nl}} \in \{0, 1\}^8$. The process of Multimodal-XAD (f_{xad}) is described as follows:

$$f_{\text{xad}}(X_i, I_i, E_i) \rightarrow (A_i, D_i^{\text{bev}}, D_i^{\text{nl}}), 1 \le i \le T, \quad (1)$$

where i and T are the index and total number of image sequences in the dataset, respectively.

We adopt EfficientNet [75] as the encoder to extract features due to its trade-off between accuracy and efficiency [76]. The encoder takes as input the image sequence X_i and outputs features F_i of images from 6 cameras. Then, the features F_i are fed into the BEV module and S-U module at the same time.

The BEV module is designed based on the Lift-Splat [39]. It contains two modules, including BEV decoder (f_{dec}) and BEV generator (f_{gen}). The function of the BEV module is to predict BEV maps D_i^{bev} of traffic scenes. The process of the BEV module is described as follows:

$$f_{\text{gen}}(f_{\text{dec}}(F_i, I_i, E_i)) \rightarrow D_i^{\text{bev}}, 1 \le i \le T, \quad (2)$$

where i and T are the index and total number of image sequences in the dataset, respectively. As the multi-class

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

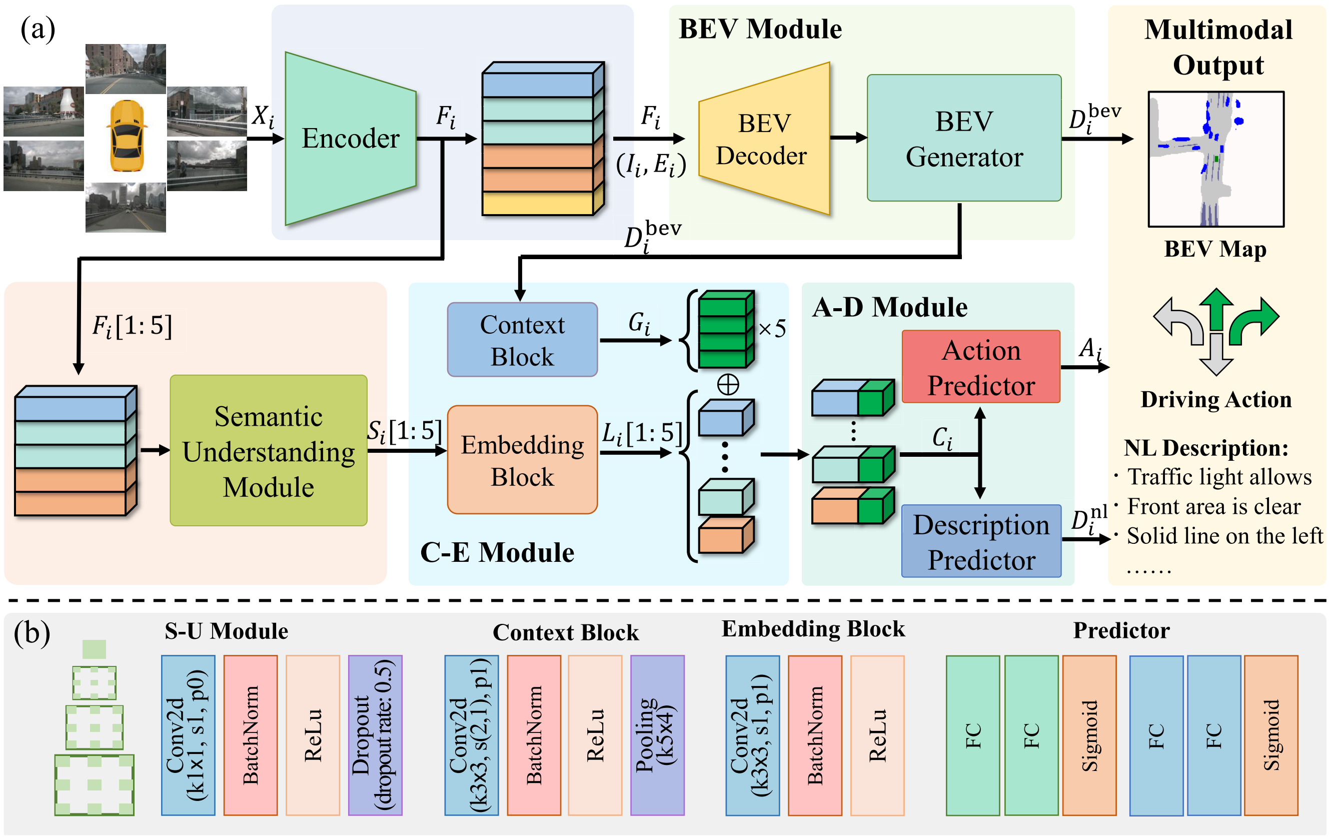


Fig. 2. The architecture of our proposed network. (a) shows the workflow of the network. The network takes as input the images along with the intrinsics and extrinsics from surrounding monocular cameras to jointly predict driving actions and multimodal environment descriptions. (b) shows the details of the S-U module, context block, embedding block and predictors. For the convolution and pooling layers, k refers to kernel size, s refers to the stride, and p refers to the padding. The figure is best viewed in color.

semantic BEV grid maps, BEV maps D_i^{bev} show the overall perception of traffic environments within 100 meter × 100 meter area. Consequently, BEV maps D_i^{bev} contain the context information of the traffic scenes.

The S-U module includes the Atrous Spatial Pyramid Pooling (ASPP, f_{aspp}) [77], convolution layer (f_{conv}, kernel size of 1×1, stride of 1, padding of 0), batch normalization layer (f_{bn}), ReLu activation layer (f_{relu}), and dropout layer (f_{drop}, dropout rate of 0.5). We refer readers to [77] for more details about the ASPP. The function of the S-U module is to enable the network to understand traffic scenes at the pixel level. It is worth noting that the inputs of the S-U module are the features $F_i[1:5]$ of images from the 5 cameras (i.e., the front, front left, front right, back left, and back right cameras). The feature $F_i[6]$ of the image from the back camera is not fed into the S-U module. The outputs of the S-U module are semantic features $S_i[1{:}5]$ of images from the 5 cameras. Unlike the BEV maps D_i^{bev} that focus on the overall perception of traffic scenes, semantic features $S_i[1{:}5]$ focus on the road details of images captured by different cameras. Consequently, semantic features $S_i[1{:}5]$ contain the local information of the traffic scenes. The process of the S-U module is described as follows:

$$f_{\text{drop}}(f_{\text{relu}}(f_{\text{bn}}(f_{\text{conv}}(f_{\text{aspp}}(F_i[1:5]))))) \to S_i[1:5], \quad 1 \le i \le T, \quad (3)$$

where i and T are the index and total number of image sequences in the dataset, respectively.

Then, BEV maps D_i^{bev} and the semantic features $S_i[1:5]$ are both fed into the C-E module. The C-E module consists of two parts: the context block (f_{cb}) and the embedding block (f_{eb}). The context block includes the convolution layer (f_{conv}, kernel size of 3×3, stride of (2, 1), padding of 1), batch normalization layer (f_{bn}), ReLu activation layer (f_{relu}), and pooling layer (f_{pool}, kernel size of 5×4). The embedding block includes the convolution layer (kernel size of 3×3, stride of 1, padding of 1), batch normalization layer (f_{bn}), and ReLu activation layer (f_{relu}). So, we have $f_{\text{cb}}(*) = f_{\text{pool}}(f_{\text{relu}}(f_{\text{bn}}(f_{\text{conv}}(*))))$, and $f_{\text{eb}}(*) = f_{\text{relu}}(f_{\text{bn}}(f_{\text{conv}}(*)))$.

In the C-E module, BEV maps D_i^{bev} and semantic features $S_i[1 : 5]$ are separately fed into the context block and the embedding block to obtain the embedded features G_i and $L_i[1 : 5]$, respectively. Then, G_i is concatenated with each $L_i[k]$ ($k \le 5$) as the concatenated features C_i. As aforementioned, BEV maps D_i^{bev} contain the context information of traffic scenes. So, considering that G_i is the embedded features for D_i^{bev}, the context information from D_i^{bev} is embedded into G_i. Similarly, the local information from $S_i[1 : 5]$ is embedded into $L_i[1 : 5]$. Eventually, by concatenating G_i and $L_i[1 : 5]$, the context and local information of traffic scenes are both encoded into the concatenated features C_i. The process of the

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

C-E module is described as follows:

$$(f_{\mathrm{cb}}(D_i^{\mathrm{bev}}) \oplus f_{\mathrm{eb}}(S_i[1:5])) \rightarrow C_i, 1 \leq i \leq T, \tag{4}$$

where i and T are the index and total number of image sequences in the dataset, respectively.

Finally, the concatenated features C_i are fed into the A-D module to predict the driving actions and the natural-language environment descriptions. The A-D module is consisted of the action predictor (f_{ap}) and the description predictor (f_{dp}). In the A-D module, the concatenated features C_i are flattened and fed into the action and description predictors to predict the driving actions and the natural-language environment descriptions. Both the action predictor and description predictor contain the two fully connected layers (f_{fc}). The Sigmoid function (f_{sgm}) is applied as the activation function after the second fully connected layer with the threshold of 0.5. So, we have $f_{\mathrm{ap}}(*) = f_{\mathrm{sgm}}(f_{\mathrm{fc}}(f_{\mathrm{fc}}(*)))$, and $f_{\mathrm{dp}}(*) = f_{\mathrm{sgm}}(f_{\mathrm{fc}}(f_{\mathrm{fc}}(*)))$. The process of the A-D module is described as follows:

$$(f_{\mathrm{ap}}(C_i), f_{\mathrm{dp}}(C_i)) \rightarrow (A_i, D_i^{\mathrm{nl}}), 1 \leq i \leq T, \tag{5}$$

where i and T are the index and total number of image sequences in the dataset, respectively.

B. Training Details

The networks are trained with the NVIDIA GeForce RTX 3090 GPU. We first pre-train the encoder and BEV module using the nuScenes dataset for 30 epochs with the batch size of 12. Then, we train Multimodal-XAD based on our proposed dataset for 60 epochs with the batch size of 8. It is worth noting that not all image sequences of the nuScenes dataset are used in the pre-training process. The image sequences of our proposed dataset are picked out and not used during pre-training. We adopt the Adam optimizer with the initial learning rate of 1×10^{-4} and weight decay of 1×10^{-8}. Our network is trained with a multi-task loss function, which is calculated as follows:

$$\mathcal{L}_{\mathrm{total}} = \lambda_1 \mathcal{L}_{\mathrm{act}} + \lambda_2 \mathcal{L}_{\mathrm{desc}}^{\mathrm{nl}} + \lambda_3 \mathcal{L}_{\mathrm{desc}}^{\mathrm{bev}}, \tag{6}$$

where $\mathcal{L}_{\mathrm{total}}$ is the total loss, $\mathcal{L}_{\mathrm{act}}$ and $\mathcal{L}_{\mathrm{desc}}^{\mathrm{nl}}$ are the binary cross entropy losses for predictions of driving actions and natural-language environment descriptions, respectively. $\mathcal{L}_{\mathrm{desc}}^{\mathrm{bev}}$ is the cross entropy loss for predictions of BEV maps. λ_1, λ_2 and λ_3 are the weight parameters that determine the relative importance between driving actions, natural-language environment descriptions and BEV maps, respectively.

IV. Experimental Results and Discussions

A. The Dataset

Tab. I shows the comparison of different explainable datasets of autonomous driving. For these available explainable datasets, they only contain natural-language explanations. To further increase the explainability, combining visual and natural-language explanations may be a more effective way to explain the outputs of autonomous driving networks. To this end, we build the *nuScenes Action and multimodal environment Descriptions* (nu-A2D) dataset that contains driving actions and multimodal environment descriptions.

TABLE I

Explainable Datasets for Autonomous Driving. The Size Refers to the Number of Explanations in the Dataset. The Action Refers to the Driving Action

Dataset	Size	Action	Explanation
BDD-X [18]	26,228	✓	Textual justification
BDD-OIA [19]	23,000	✓	Natural-language reasons
BDD-AD [26]	10,000	✓	Natural-language descriptions
HDD [73]	47,533	✓	Textual causal reasoning
PSI [74]	11,902	✓	Text-based reasons
nu-A2D	12,000	✓	Multimodal environment descriptions

TABLE II

The Categories of Driving Actions and Natural-Language Environment Descriptions (Labelled as NL Description) in Our Proposed Nu-A2D Dataset. The Ratio Refers to the Percentage of Each Category in the Dataset

Annotation	Category	Ratio (%)
Driving Action	Move forward	80.76
	Stop/slow down	19.23
	Turn left/change to left lane	29.43
	Turn right/change to right lane	38.07
NL Description	Traffic light allows	84.14
	Front area is clear	89.64
	Solid line on the left	27.61
	Solid line on the right	23.74
	Front left area is clear	41.57
	Back left area is clear	42.91
	Front right area is clear	46.93
	Back right area is clear	50.54

In the nu-A2D dataset, 12, 000 image sequences are selected from the nuScenes [67] dataset. Each image sequence contains 6 images from surrounding cameras along with the ground truth of driving actions, natural-language environment descriptions and BEV maps for traffic scenes. The ground truth of driving actions and natural-language environment descriptions are manually labelled by ourselves. For each image sequence, we need to observe 6 images from surrounding cameras to determine and label the appropriate driving actions and natural-language environment descriptions. Specifically, 4 categories of driving actions and 8 categories of natural-language environment descriptions are adopted for the nu-A2D dataset. Tab. II summarizes the categories and ratios of driving actions and natural-language environment descriptions. The ground truth of the BEV map is obtained by projecting 3D bounding boxes of objects into the BEV plane and transforming map layers from the nuScenes map into the ego-vehicle frame.

B. Evaluation Metrics

To evaluate the prediction performance of BEV maps, the Intersection over Union (IoU) values for road, vehicle and road/lane divider are calculated. Take the semantic class of road for example, its IoU value is calculated as follows:

$$\mathrm{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} \times 100\%, \tag{7}$$

where the area of intersection refers to the area where both the predicted BEV map and the ground truth have the semantic

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

TABLE III

COMPARATIVE RESULTS OF THE PREDICTION PERFORMANCE OF DRIVING ACTIONS FOR DIFFERENT NETWORKS. LABEL F DENOTES "MOVE FORWARD", LABEL S DENOTES "STOP/SLOW DOWN", LABEL L DENOTES "TURN LEFT/CHANGE TO LEFT LANE" AND LABEL R DENOTES "TURN RIGHT/CHANGE TO RIGHT LANE". THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Networks	F	S	L	R	$F1_{\mathrm{m}}^{\mathrm{act}}$	$F1_{\mathrm{oval}}^{\mathrm{act}}$
Decision Model [25]	0.927	0.621	0.805	0.805	0.790	0.863
VPN [42]	0.916	0.280	0.806	0.866	0.717	0.859
CVT [46]	0.936	0.743	0.835	0.880	0.849	0.898
Multimodal-XAD	0.959	0.798	0.847	0.875	**0.870**	**0.913**

class of road. The area of union refers to the total area where either the predicted BEV map or the ground truth has the semantic class of road. The mean IoU (mIoU) is also calculated to show the average IoU value for all semantic classes.

To evaluate the prediction performance of driving actions and natural-language environment descriptions, the standard F1 score metric is employed in this work. Specifically, the overall and mean F1 score are used. The overall F1 score is calculated as:

$$F1_{\mathrm{oval}}^{\mathrm{act}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{F1}(A_i, \hat{A}_i), \tag{8}$$

$$F1_{\mathrm{oval}}^{\mathrm{desc}} = \frac{1}{M}\sum_{j=1}^{M}\mathrm{F1}(D_j, \hat{D}_j), \tag{9}$$

where $F1_{\mathrm{oval}}^{\mathrm{act}}$ and $F1_{\mathrm{oval}}^{\mathrm{desc}}$ are the overall F1 scores for the predictions of driving actions and natural-language environment descriptions, respectively. A_i and $\hat{A}_i$ are the prediction and ground truth for driving actions, respectively. D_j and $\hat{D}_j$ are the prediction and ground truth for natural-language environment descriptions, respectively. N and M are the numbers of the driving action predictions and natural-language environment description predictions, respectively.

Given that the nu-A2D dataset is imbalanced, in which the ratio between each category of driving actions and natural-language environment descriptions is different (see detail in Tab. II), we also calculate the mean F1 score for predictions of driving actions and natural-language environment descriptions. The mean F1 score for the predictions of driving actions is calculated as follows:

$$\begin{aligned} F1_{\mathrm{m}}^{\mathrm{act}} = \frac{1}{4}\Big(&\frac{1}{N^{\mathrm{f}}}\sum_{i=1}^{N^{\mathrm{f}}}\mathrm{F1}(A_i^{\mathrm{f}}, \hat{A}_i^{\mathrm{f}}) + \frac{1}{N^{\mathrm{s}}}\sum_{j=1}^{N^{\mathrm{s}}}\mathrm{F1}(A_j^{\mathrm{s}}, \hat{A}_j^{\mathrm{s}}) \\ &+\frac{1}{N^{\mathrm{l}}}\sum_{k=1}^{N^{\mathrm{l}}}\mathrm{F1}(A_k^{\mathrm{l}}, \hat{A}_k^{\mathrm{l}}) + \frac{1}{N^{\mathrm{r}}}\sum_{p=1}^{N^{\mathrm{r}}}\mathrm{F1}(A_p^{\mathrm{r}}, \hat{A}_p^{\mathrm{r}})\Big), \end{aligned} \tag{10}$$

where $F1_{\mathrm{m}}^{\mathrm{act}}$ is the mean F1 score for predictions of driving actions. N^{f}, N^{s}, N^{l}, and N^{r} are the numbers of the predictions for "move forward", "stop/slow down", "turn left/change to left lane" and "turn right/change to right lane", respectively. A_i^{f} and $\hat{A}_i^{\mathrm{f}}$ represent the prediction and ground truth for "move forward", respectively. The same naming rules are used for A_j^{s}, $\hat{A}_j^{\mathrm{s}}$, A_k^{l}, $\hat{A}_k^{\mathrm{l}}$, A_p^{r} and $\hat{A}_p^{\mathrm{r}}$.

The mean F1 score for the predictions of natural-language environment descriptions is calculated as follows:

$$F1_{\mathrm{m}}^{\mathrm{desc}} = \frac{1}{8}\sum_{e=1}^{8}\Big(\frac{1}{M^e}\sum_{i=1}^{M^e}\mathrm{F1}(D_i^e, \hat{D}_i^e)\Big), \tag{11}$$

where $F1_{\mathrm{m}}^{\mathrm{desc}}$ is the mean F1 score for predictions of natural-language environment descriptions. There are 8 categories of natural-language environment descriptions, so e ranges from 1 to 8. M^e represents the number of the predictions of each natural-language environment description category. D_i^e and $\hat{D}_i^e$ represent the prediction and ground truth for each natural-language environment description category, respectively.

C. Comparative Results

Tab. III and Tab. IV show the comparative results of the prediction performance of driving actions and multimodal environment descriptions, respectively. Decision Model [25] is trained on the A2D dataset (without BEV maps) to jointly predict the driving actions and the natural-language environment descriptions. Both VPN [42] and CVT [46] are modified and trained on the nu-A2D dataset to jointly predict the driving actions and multimodal environment descriptions. In VPN and CVT, the driving action and natural-language environment description are predicted only based on the context information of traffic scenes. To ensure the fairness of comparison, VPN, CVT and Multimodal-XAD are all pre-trained with the same epochs on the nuScenes dataset before training on the nu-A2D dataset.

As shown in Tab. III, both the $F1_{\mathrm{m}}^{\mathrm{act}}$ and $F1_{\mathrm{oval}}^{\mathrm{act}}$ of Multimodal-XAD are higher than those of the other networks. Specifically, the $F1_{\mathrm{m}}^{\mathrm{act}}$ of Multimodal-XAD is about 10%, 21% and 2% higher than those of Decision Model, VPN and CVT, respectively. The $F1_{\mathrm{oval}}^{\mathrm{act}}$ of Multimodal-XAD is about 6%, 6% and 2% higher than those of Decision Model, VPN and CVT, respectively. These comparative results show that Multimodal-XAD presents better prediction performance of driving action than the other networks. The better prediction performance of driving actions would lead to higher safety for autonomous driving. Tab. IV shows the comparative results of the prediction performance of multimodal environment descriptions for different networks. The $F1_{\mathrm{m}}^{\mathrm{desc}}$ and $F1_{\mathrm{oval}}^{\mathrm{desc}}$ of Multimodal-XAD and CVT are very close and higher than those of Decision Model and VPN. Specifically, the $F1_{\mathrm{m}}^{\mathrm{desc}}$ of Multimodal-XAD is about 4% and 1% higher than those of Decision Model and VPN, respectively. The $F1_{\mathrm{oval}}^{\mathrm{desc}}$ of

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

COMPARATIVE RESULTS OF THE PREDICTION PERFORMANCE OF MULTIMODAL ENVIRONMENT DESCRIPTIONS FOR DIFFERENT NETWORKS. THE NATURAL-LANGUAGE ENVIRONMENT DESCRIPTION IS LABELLED AS NL DESCRIPTION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Descriptions	Categories	F1 Score / IoU (%)			
		Decision Model [25]	VPN [42]	CVT [46]	Multimodal-XAD
NL Description	Traffic light allows	0.934	0.900	0.928	0.952
	Front area is clear	0.955	0.958	0.970	0.973
	Solid line on the left	0.892	0.907	0.893	0.886
	Solid line on the right	0.876	0.920	0.784	0.811
	Front left area is clear	0.794	0.869	0.862	0.893
	Back left area is clear	–	0.730	0.891	0.899
	Front right area is clear	0.623	0.820	0.877	0.885
	Back right area is clear	–	0.844	0.810	0.750
	$F1^{\mathrm{desc}}_{\mathrm{m}}$	0.846	0.869	0.877	**0.881**
	$F1^{\mathrm{desc}}_{\mathrm{oval}}$	0.857	0.859	0.893	**0.897**
BEV Map	Road	–	60.6	57.5	59.5
	Vehicle	–	25.1	22.8	25.7
	Road/lane divider	–	21.8	26.0	31.0
	mIoU	–	35.8	35.4	**38.7**

Multimodal-XAD is about 5% and 4% higher than those of Decision Model and VPN, respectively. For the mIoU of BEV maps, Multimodal-XAD is about 8% and 9% higher than those of VPN and CVT, respectively. The better prediction performance of multimodal environment descriptions could lead to more effective and accurate explanations for driving actions.

As aforementioned, the segmentation performance of vision-based BEV perception is limited and can further influence the prediction performance of downstream tasks. Unlike VPN and CVT, Multimodal-XAD predicts driving actions and natural-language environment descriptions based on both the context information from BEV perception and local information from semantic perception. So, the utilization of the local information may alleviate the error accumulation. We believe that this may be the possible reason why the prediction performance of Multimodal-XAD is better than VPN and CVT. On the other hand, Decision Model predicts the driving actions and natural-language environment descriptions only based on the local information. The absence of context information hinders Decision Model from attaining a more comprehensive understanding of traffic scenes. This may be the reason the prediction performance of Multimodal-XAD is better than Decision Model.

To evaluate the computational complexity of different networks, three key metrics are used, including the number of parameters (Param), Floating Point Operations (FLOPs), and Frames Per Second (FPS) for the inference. As shown in Tab. V, the Param of Decision Model and CVT are very close, which are both lower than Multimodal-XAD. However, Multimodal-XAD has the lowest FLOPs among these networks, resulting in the highest FPS for inference.

TABLE V

COMPUTATIONAL COMPLEXITY FOR DIFFERENT NETWORKS ON THE NU-A2D DATASET. THE INFERENCE SPEED IS TESTED USING AN NVIDIA GEFORCE RTX 3060 GPU

Configuration	Param	FLOPs	FPS
Decision Model [25]	6.96M	82.91G	18.06
VPN [42]	82.92M	263.79G	13.91
CVT [46]	6.82M	69.40G	19.96
Multimodal-XAD	14.78M	41.57G	21.14

Fig. 3 shows sample qualitative results for different networks. Complex traffic scenes are chosen to validate the prediction performance and generalization capability of our proposed network. Unlike the other networks, all driving action predictions and most natural-language environment description predictions of Multimodal-XAD are the same as ground truth, which demonstrates that Multimodal-XAD is able to perceive and understand traffic scenes and predict the correct driving actions and natural-language environment descriptions. For the visualization results of BEV maps, Multimodal-XAD generates more precise and clear BEV maps than the other networks. Specifically, we can see that Multimodal-XAD is more sensitive to vehicles and road/lane dividers compared to the other networks, which is critical for safe navigation.

To further validate the prediction performance of our Multimodal-XAD, we have tested it on the BDD-OIA [19] dataset. In the BDD-OIA dataset, each image is labelled with 4 driving actions and 21 natural-language reasons. Our Multimodal-XAD is modified to jointly predict the driving actions and corresponding reasons. Tab. VI shows the comparative results of the prediction performance for different networks on the BDD-OIA dataset. In Tab. VI, $F1^{\mathrm{act}}_{\mathrm{m}}$ and $F1^{\mathrm{act}}_{\mathrm{oval}}$ refer to the mean and overall F1 scores for predictions of driving actions, respectively. $F1^{\mathrm{rea}}_{\mathrm{m}}$ and $F1^{\mathrm{rea}}_{\mathrm{oval}}$ refer to the mean and overall F1 scores for predictions of natural-language reasons, respectively. For $F1^{\mathrm{act}}_{\mathrm{m}}$, $F1^{\mathrm{act}}_{\mathrm{oval}}$ and $F1^{\mathrm{rea}}_{\mathrm{m}}$, our Multimodal-XAD is the highest among all these networks. For $F1^{\mathrm{rea}}_{\mathrm{oval}}$, Multimodal-XAD and Interrelation Model are very close and higher than the other networks. These comparative results demonstrate that our proposed Multimodal-XAD has

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

Comparative Results of the Prediction Performance for Different Networks on the BDD-OIA [19] Dataset. Label F Denotes “Move Forward”, Label S Denotes “stop/Slow Down”, Label L Denotes “Turn Left/Change to Left Lane” and Label R Denotes “Turn Right/Change to Right Lane”. $F1_{\mathrm{M}}^{\mathrm{REA}}$ and $F1_{\mathrm{OVAL}}^{\mathrm{REA}}$ Refer to the Mean and Overall F1 Scores for Predictions of Natural-Language Reasons, Respectively. The Best Results Are Highlighted in Bold Font

Methods	F	S	L	R	$F1_{\mathrm{m}}^{\mathrm{act}}$	$F1_{\mathrm{oval}}^{\mathrm{act}}$	$F1_{\mathrm{m}}^{\mathrm{rea}}$	$F1_{\mathrm{oval}}^{\mathrm{rea}}$
OIA [19]	0.829	0.781	0.630	0.634	0.718	0.734	0.208	0.422
CBM-AUC [20]	0.803	0.751	0.551	0.525	0.658	0.704	0.342	0.522
C-SENN [21]	0.772	0.744	0.469	0.486	0.618	–	–	–
CBM [22]	0.795	0.732	0.483	0.431	0.610	0.661	0.292	0.412
Interrelation Model [23]	0.802	0.753	0.619	0.625	0.701	0.722	0.335	**0.537**
Multimodal-XAD	0.822	0.789	0.638	0.641	**0.723**	**0.743**	**0.360**	0.535

better prediction performance in terms of both driving actions and the corresponding natural-language reason than the other networks.

D. Ablation Study

In the ablation study, we first investigate the influence of the context and local information on the prediction performance of driving actions and multimodal environment descriptions. Tab. VII shows the ablation study results of the prediction performance of driving actions for different networks. For the `No Context` network, the context information is not fed into the C-E Module to concatenate with the local information. The driving actions and natural-language environment descriptions of the `No Context` network are predicted only based on the local information from the semantic perception. On the opposite, for the `No Local` network, the local information is not fed into the C-E Module to concatenate with the context information. The driving actions and natural-language environment descriptions of the `No Local` network are predicted only based on the context information from the BEV perception. As shown in Tab. VII, both the $F1_{\mathrm{m}}^{\mathrm{act}}$ and $F1_{\mathrm{oval}}^{\mathrm{act}}$ of Multimodal-XAD are higher than those of the `No Context` and `No Local` networks. These results validate that the combination of context and local information would improve the prediction performance of driving actions.

Tab. VIII shows the ablation study results of the prediction performance of multimodal environment descriptions for different networks. As shown in Tab. VIII, both the $F1_{\mathrm{m}}^{\mathrm{desc}}$ and $F1_{\mathrm{oval}}^{\mathrm{desc}}$ of Multimodal-XAD are higher than those of the `No Context` and `No Local` networks. These results validate that the combination of the context and local information could improve the prediction performance of natural-language environment description. For the prediction performance of BEV maps, the mIoU of Multimodal-XAD, `No Context` and `No Local` networks are at the same level. This indicates that using context and local information to predict driving actions and natural-language environment descriptions has little impact on the prediction of BEV maps.

The influence of the natural-language environment description on the prediction performance of driving actions and BEV maps is also investigated. As shown in Tab. VII, both the $F1_{\mathrm{m}}^{\mathrm{act}}$ and $F1_{\mathrm{oval}}^{\mathrm{act}}$ of Multimodal-XAD are higher than those of the network without natural-language environment description (labelled as `No NLD`). This result shows that the natural-language environment description would improve the prediction performance of driving actions. For the prediction performance of BEV maps (as shown in Tab. VIII), the mIoU of Multimodal-XAD is lower than the `No NLD` network, which shows that the natural-language environment description has no positive influence on the prediction of BEV maps.

Fig. 3 also shows sample qualitative results for the `No Context`, `No Local` and `No NLD` networks. For the `No Context` and `No Local` networks, the prediction performance of driving actions and natural-language environment descriptions are worse than Multimodal-XAD, indicating that the combination of context and local information could improve the prediction performance of driving actions natural-language environment descriptions. For the `No NLD` network, the explainability is lower compared to Multimodal-XAD due to the lack of natural-language environment descriptions.

In this section, we further investigate the influence of different encoders on the prediction performance of Multimodal-XAD. Fig. 4 shows the ablation study results of the prediction performance for Multimodal-XAD networks with different EfficientNet variants, including Efficient-B0 to Efficient-B7. The left figure shows the F1 scores of the driving action and natural-language environment description predictions of Multimodal-XAD networks with different encoders. As shown in the left figure of Fig. 4, both the $F1_{\mathrm{m}}^{\mathrm{act}}$ and $F1_{\mathrm{oval}}^{\mathrm{act}}$ of Multimodal-XAD with the Efficient-B4 has the highest F1 score. This result shows that the driving action prediction performance of Multimodal-XAD with the Efficient-B4 is the best among all the networks with different encoders. For the prediction performance of natural-language environment description, both the $F1_{\mathrm{m}}^{\mathrm{desc}}$ and $F1_{\mathrm{oval}}^{\mathrm{desc}}$ of Multimodal-XAD with Efficient-B6 is the highest. The $F1_{\mathrm{m}}^{\mathrm{desc}}$ and $F1_{\mathrm{oval}}^{\mathrm{desc}}$ of Multimodal-XAD with Efficient-B4 is the fourth highest and second highest, respectively.

The right figure of Fig. 4 shows the IoU of different classes of BEV map predictions of Multimodal-XAD with different encoders. As shown in the right figure of Fig. 4, the rising trend of the prediction performance of BEV maps can be seen when increasing the complexity of EfficientNet from B0 to B7. Therefore, to trade off prediction performance and computation cost, we choose EfficientNet-B4 as the default encoder of our Multimodal-XAD.

The influence of the relative importance between driving actions, natural-language environment descriptions and BEV

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

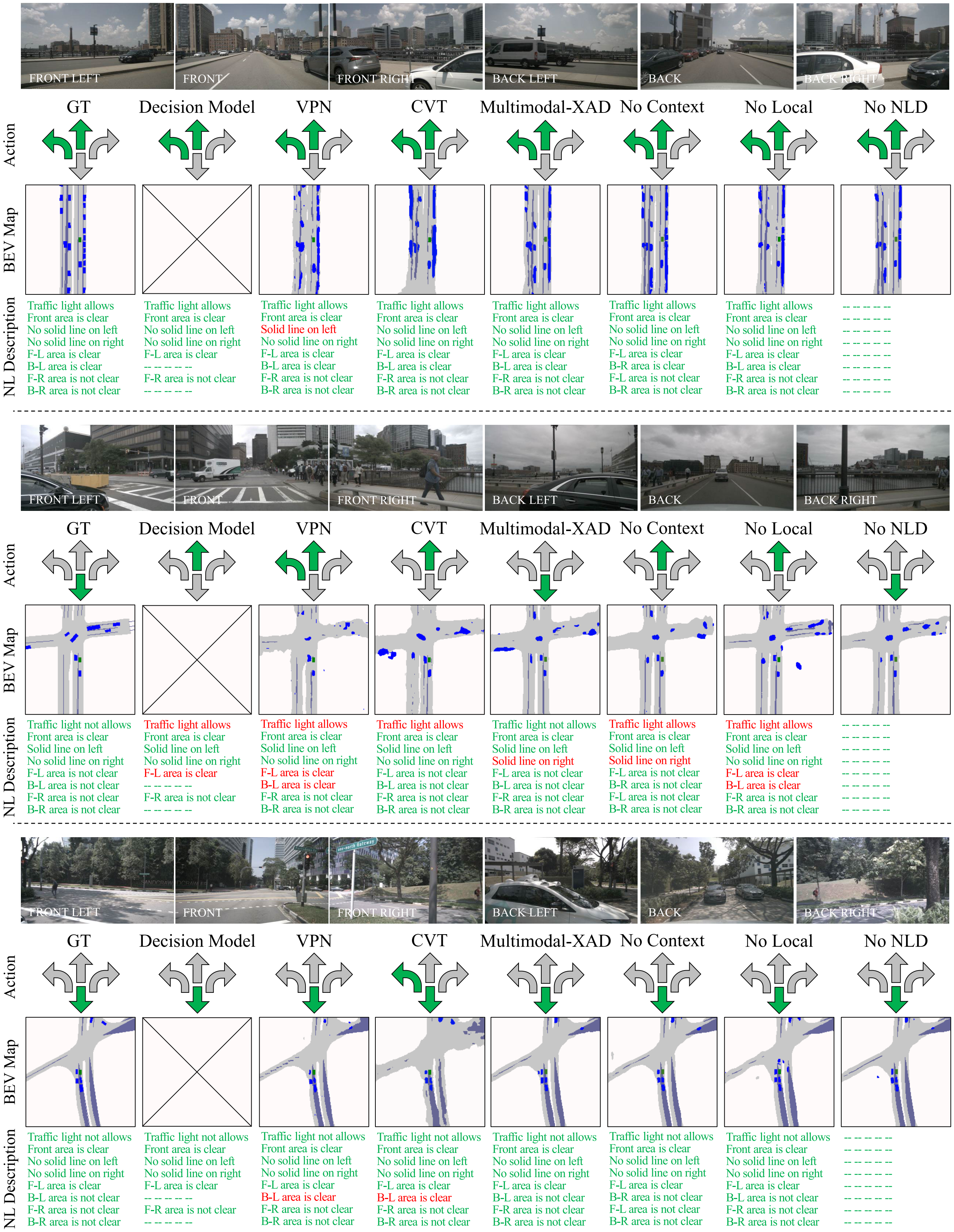


Fig. 3. Sample qualitative results of predictions of driving actions and multimodal environment descriptions for different networks. The label "GT" denotes the ground truth for driving action and multimodal environment descriptions. For the predictions of natural-language environment description (labelled as NL description), the green one denotes correct prediction, and the red one denotes wrong predictions. The figure is best viewed in color.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

ABLATION STUDY RESULTS OF THE PREDICTION PERFORMANCE OF DRIVING ACTIONS FOR DIFFERENT NETWORKS. LABEL F DENOTES "MOVE FORWARD", LABEL S DENOTES "STOP/SLOW DOWN", LABEL L DENOTES "TURN LEFT/CHANGE TO LEFT LANE" AND LABEL R DENOTES "TURN RIGHT/CHANGE TO RIGHT LANE". THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Networks	F	S	L	R	$F1_{\text{m}}^{\text{act}}$	$F1_{\text{oval}}^{\text{act}}$
Multimodal-XAD	0.959	0.798	0.847	0.875	**0.870**	**0.913**
No Context	0.932	0.584	0.827	0.847	0.798	0.879
No Local	0.940	0.618	0.835	0.848	0.810	0.887
No NLD	0.938	0.655	0.820	0.868	0.820	0.887

TABLE VIII

ABLATION STUDY RESULTS OF THE PREDICTION PERFORMANCE OF MULTIMODAL ENVIRONMENT DESCRIPTIONS FOR DIFFERENT NETWORKS. THE NATURAL-LANGUAGE ENVIRONMENT DESCRIPTION IS LABELLED AS NL DESCRIPTION. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

Descriptions	Categories	F1 Score / IoU (%)			
		Multimodal-XAD	No Context	No Local	No NLD
NL Description	Traffic light allows	0.952	0.933	0.928	–
	Front area is clear	0.973	0.969	0.966	–
	Solid line on the left	0.886	0.860	0.826	–
	Solid line on the right	0.811	0.802	0.835	–
	Front left area is clear	0.893	0.901	0.882	–
	Back left area is clear	0.899	0.878	0.835	–
	Front right area is clear	0.885	0.856	0.820	–
	Back right area is clear	0.750	0.752	0.782	–
	$F1_{\text{m}}^{\text{desc}}$	**0.881**	0.869	0.859	–
	$F1_{\text{oval}}^{\text{desc}}$	**0.897**	0.884	0.869	–
BEV Map	Road	59.5	59.2	60.7	61.3
	Vehicle	25.7	25.8	26.2	27.6
	Road/lane divider	31.0	31.1	29.0	31.6
	mIoU	38.7	38.7	38.6	**40.2**

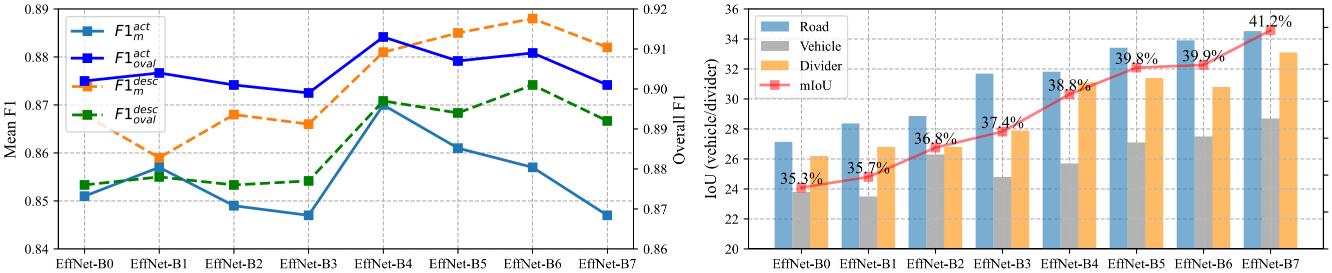


Fig. 4. Ablation study results of the prediction performance of Multimodal-XAD with different encoders of the EfficientNet family. The left figure shows the F1 scores of the driving action and natural-language environment description predictions. The right figure shows the IoU of predictions of BEV maps. EffNet is the short for EfficientNet. The figure is best viewed in color.

TABLE IX

THE ABLATION STUDY RESULTS OF PREDICTION PERFORMANCE FOR MULTIMODAL-XAD NETWORKS WITH DIFFERENT RELATIVE IMPORTANCE BETWEEN DRIVING ACTIONS, NATURAL-LANGUAGE ENVIRONMENT DESCRIPTIONS AND BEV MAPS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT

$\lambda_1 : \lambda_2 : \lambda_3$	$F1_{\text{m}}^{\text{act}}$	$F1_{\text{oval}}^{\text{act}}$	mIoU (%)	$F1_{\text{m}}^{\text{desc}}$	$F1_{\text{oval}}^{\text{desc}}$
1:1:1	0.870	0.913	38.7	0.881	0.897
2:1:1	**0.873**	**0.914**	38.0	0.863	0.875
1:2:1	0.855	0.903	37.8	**0.895**	**0.906**
1:1:2	0.828	0.892	**40.3**	0.878	0.892

maps on the prediction performance of Multimodal-XAD is also investigated in the ablation study. The relative importance is determined by λ_1, λ_2 and λ_3 on the loss function (6). Here, we have tested four different configurations of relative importance. As shown in Tab. IX, Multimodal-XAD with higher importance of driving actions ($\lambda_1 = 2$) has the best prediction performance of driving actions. Similarly, Multimodal-XAD with higher importance of natural-language environment descriptions ($\lambda_2 = 2$) and BEV maps ($\lambda_3 = 2$) exhibit the best prediction performance in natural-language descriptions and BEV maps, respectively. To achieve a more balanced performance of driving actions, natural-language environment descriptions and BEV maps, we chose $\lambda_1 : \lambda_2 : \lambda_3 = 1 : 1 : 1$ as the default configuration for our Multimodal-XAD.

E. Limitations

Despite the superiority of our proposed Multimodal-XAD, there are still some limitations. Firstly, the prediction

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

performance of the multimodal environment description is evaluated by calculating the IoU of the BEV map and the F1 score of the natural-language environment description, respectively. To better measure the explainability of Multimodal-XAD, a new metric that could comprehensively evaluate the multimodal environment description should be exploited. Secondly, the driving action is coupled with the BEV perception by using the context information to predict the driving action in Multimodal-XAD. However, we believe that driving action and BEV perception could be more tightly coupled to improve the prediction performance of driving actions. For example, we can apply the generative adversarial network (GAN), in which the driving action predictor is served as the generator and the BEV module is served as the discriminator.

V. Conclusion and Future Work

To improve the safety and explainability of deep learning-based autonomous driving, we proposed an explainable autonomous driving network that jointly predicts the driving actions and the multimodal environment descriptions of traffic scenes, including BEV maps and natural-language environment descriptions. In the proposed network, both the context information from BEV perception and the local information from semantic perception are considered before predicting the driving actions and natural-language environment descriptions. A new dataset containing $12,000$ image sequences is released. Each image sequence contains 6 frames from surrounding visual cameras, as well as hand-labelled ground truth for driving actions and multimodal environment descriptions. The experimental results show that the combination of context information and local information improves the prediction performance of both driving actions and environment descriptions.

Regarding future work, one promising research direction could be the utilization of temporal information in Multimodal-XAD. At the current stage, the input of our proposed network is the image sequence from surrounding cameras. In other words, no temporal information is considered before predicting driving actions and multimodal environment descriptions. However, the real-world traffic environment is dynamic and interactive. So, we believe that utilizing the consecutive image sequence that contains temporal information of traffic scenes may improve the safety and explainability of Multimodal-XAD.

References

[1] S. Teng et al., "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 6, pp. 3692–3711, Jun. 2023.

[2] P. Cai, S. Wang, Y. Sun, and M. Liu, "Probabilistic end-to-end vehicle navigation in complex dynamic environments with multimodal sensor fusion," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4218–4224, Jul. 2020.

[3] L. Chen, S. Lin, X. Lu, D. Cao, and F. Y. Wang, "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021.

[4] P. Cai, X. Mei, L. Tai, Y. Sun, and M. Liu, "High-speed autonomous drifting with deep reinforcement learning," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1247–1254, Apr. 2020.

[5] H. Xu, H. Liu, S. Huang, and Y. Sun, "C2L-PR: Cross-modal camera-to-LiDAR place recognition via modality alignment and orientation voting," *IEEE Trans. Intell. Veh.*, early access, Jul. 4, 2024, doi: 10.1109/TIV.2024.3423392.

[6] Z. Sheng, Y. Xu, S. Xue, and D. Li, "Graph-based spatial–temporal convolutional network for vehicle trajectory prediction in autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17654–17665, Oct. 2022.

[7] Z. Feng, Y. Guo, and Y. Sun, "Segmentation of road negative obstacles based on dual semantic-feature complementary fusion for autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 4, pp. 4687–4697, Apr. 2024.

[8] P. Cai, Y. Sun, H. Wang, and M. Liu, "VTGNet: A vision-based trajectory generation network for autonomous vehicles in urban environments," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 3, pp. 419–429, Sep. 2021.

[9] S. Teng, L. Chen, Y. Ai, Y. Zhou, Z. Xuanyuan, and X. Hu, "Hierarchical interpretable imitation learning for end-to-end autonomous driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 673–683, Jan. 2023.

[10] P. Cai, H. Wang, Y. Sun, and M. Liu, "DQ-GAT: Towards safe and efficient autonomous driving with deep Q-learning and graph attention networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21102–21112, Nov. 2022.

[11] Y. Feng and Y. Sun, "PolarPoint-BEV: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving," *IEEE Trans. Intell. Veh.*, early access, Feb. 1, 2024, doi: 10.1109/TIV.2024.3361093.

[12] M. Bojarski et al., "Explaining how a deep neural network trained with end-to-end learning steers a car," 2017, *arXiv:1704.07911*.

[13] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2942–2950.

[14] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "PlanT: Explainable planning transformers via object-level representations," in *Proc. Conf. Robotic Learn. (CoRL)*, 2022, pp. 459–470.

[15] H. Wang, P. Cai, Y. Sun, L. Wang, and M. Liu, "Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 13731–13737.

[16] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "ST-P3: End-to-end vision-based autonomous driving via spatial–temporal feature learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 533–549.

[17] J. Chen, S. E. Li, and M. Tomizuka, "Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5068–5078, Jun. 2022.

[18] J. Kim et al., "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 563–578.

[19] Y. Xu et al., "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9523–9532.

[20] Y. Sawada and K. Nakamura, "Concept bottleneck model with additional unsupervised concepts," *IEEE Access*, vol. 10, pp. 41758–41765, 2022.

[21] Y. Sawada and K. Nakamura, "C-SENN: Contrastive self-explaining neural network," 2022, *arXiv:2206.09575*.

[22] P. W. Koh et al., "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.

[23] Z. Zhang, R. Tian, R. Sherony, J. Domeyer, and Z. Ding, "Attention-based interrelation modeling for explainable automated driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 2, pp. 1564–1573, Feb. 2023.

[24] H. Ben-Younes, É. Zablocki, P. Pérez, and M. Cord, "Driving behavior explanation with multi-level fusion," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108421.

[25] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering counterfactual explanations with semantics," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 387–403.

[26] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9780–9791, Sep. 2023.

[27] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," 2013, *arXiv:1312.6034*.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

[29] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 839–847.

[30] Y. Ma et al., "Vision-centric BEV perception: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 9, 2024, doi: 10.1109/TPAMI.2024.3449912.

[31] J. Fang, F. Wang, J. Xue, and T.-S. Chua, "Behavioral intention prediction in driving scenes: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 8334–8355, Aug. 2024.

[32] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: Exploiting radar for robust perception of dynamic objects," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 496–512.

[33] S. T. Isele, F. Klein, M. Brosowsky, and J. M. Zöllner, "Learning semantics on radar point-clouds," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 810–817.

[34] R. Van Kempen, B. Lampe, T. Woopen, and L. Eckstein, "A simulation-based end-to-end learning framework for evidential occupancy grid mapping," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jul. 2021, pp. 934–939.

[35] Z. Lin, Y. Wang, S. Qi, N. Dong, and M.-H. Yang, "BEV-MAE: Bird's eye view masked autoencoders for point cloud pre-training in autonomous driving scenarios," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 3531–3539.

[36] Y. Kim and D. Kum, "Deep learning based vehicle position and orientation estimation via inverse perspective mapping image," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 317–323.

[37] L. Reiher, B. Lampe, and L. Eckstein, "A Sim2Real deep learning approach for the transformation of images from multiple vehicle-mounted cameras to a semantically segmented image in Bird's eye view," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.

[38] S. Gao, Q. Wang, and Y. Sun, "S2G2: Semi-supervised semantic bird-eye-view grid-map generation using a monocular camera for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11974–11981, Oct. 2022.

[39] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 194–210.

[40] A. Hu et al., "FIERY: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15273–15282.

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

[42] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jul. 2020.

[43] N. Hendy et al., "FISHING net: Future inference of semantic heatmaps in grids," 2020, *arXiv:2006.09917*.

[44] T. Roddick and R. Cipolla, "Predicting semantic map representations from images using pyramid occupancy networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 11138–11147.

[45] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2022, pp. 180–191.

[46] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13760–13769.

[47] Y. Liu et al., "PETRv2: A unified framework for 3D perception from multi-camera images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3262–3272.

[48] S. Chen, Y. Ma, Y. Qiao, and Y. Wang, "M-BEV: Masked bev perception for robust autonomous driving," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 1183–1191.

[49] T. Liang et al., "BEVFusion: A simple and robust LiDAR-camera fusion framework," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10421–10434.

[50] Y. Man, L.-Y. Gui, and Y.-X. Wang, "BEV-guided multi-modality fusion for driving perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 21960–21969.

[51] A. W. Harley, Z. Fang, J. Li, R. Ambrus, and K. Fragkiadaki, "Simple-BEV: What really matters for multi-sensor BEV perception?" in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2759–2765.

[52] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2774–2781.

[53] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[54] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, Jan. 2022.

[55] H. Li, H. K. Chu, and Y. Sun, "Temporal consistency for RGB-thermal data-based semantic scene understanding," *IEEE Robot. Autom. Lett.*, early access, Sep. 10, 2024, doi: 10.1109/LRA.2024.3458594.

[56] J. Huang et al., "RoadFormer+: Delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion," *IEEE Trans. Intell. Veh.*, early access, Aug. 22, 2024, doi: 10.1109/TIV.2024.3448251.

[57] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7077–7087.

[58] J. Fang, L.-L. Li, K. Yang, Z. Zheng, J. Xue, and T.-S. Chua, "Cognitive accident prediction in driving scenes: A multimodality benchmark," 2022, *arXiv:2212.09381*.

[59] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[60] L. Chen et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2024, pp. 14093–14100.

[61] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robot. Autom. Lett.*, vol. 9, no. 10, pp. 8186–8193, Oct. 2024.

[62] C. Sima et al., "DriveLM: Driving with graph visual question answering," 2023, *arXiv:2312.14150*.

[63] D. Fu et al., "Drive like a human: Rethinking autonomous driving with large language models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 910–919.

[64] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[65] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[66] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2702–2719, Oct. 2020.

[67] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[68] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2636–2645.

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

[70] G. Rong et al., "LGSVL simulator: A high fidelity simulator for autonomous driving," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.

[71] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "DriveDreamer: Towards Real-world-driven world models for autonomous driving," 2023, *arXiv:2309.09777*.

[72] Y. Guan et al., "World models for autonomous driving: An initial survey," *IEEE Trans. Intell. Veh.*, early access, May 8, 2024, doi: 10.1109/TIV.2024.3398357.

[73] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7699–7707.

[74] T. Chen et al., "PSI: A pedestrian behavior dataset for socially intelligent autonomous car," 2021, *arXiv:2112.02604*.

[75] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[76] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-Maadeed, "Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches," 2022, *arXiv:2206.08016*.

[77] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

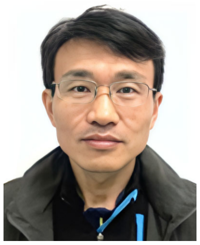
Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.

Yuchao Feng (Student Member, IEEE) received the bachelor's degree from Taiyuan University of Technology, Taiyuan, China, in 2016, and the master's degree from the University of Science and Technology of China, Hefei, China, in 2019. He is currently pursuing the Ph.D. degree with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong.

His current research interests include autonomous driving, explainable artificial intelligence, and deep learning.

Zhen Feng (Member, IEEE) received the B.S., M.S., and first Ph.D. degrees from Harbin Institute of Technology, Harbin, China, in 2019, 2017, and 2023, respectively, and the second Ph.D. degree from The Hong Kong Polytechnic University, Hung Hom, Hong Kong, in 2024.

He is currently a Post-Doctoral Fellow with the Department of Mechanical Engineering, The Hong Kong Polytechnic University. His current research interests include semantic segmentation, computer vision, autonomous driving, and deep learning.



Wei Hua received the Ph.D. degree in applied mathematics from Zhejiang University. He is currently a Senior Research Expert with Zhejiang Laboratory, Hangzhou, China. His current research interests include autonomous driving, intelligent simulation, digital twin, reinforcement learning, and AI-based algorithms.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from Hefei University of Technology, Hefei, China, in 2009, the master's degree from the University of Science and Technology of China, Hefei, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2017.

He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His current research interests include robotics and AI, autonomous driving, mobile robots, and autonomous navigation.

Dr. Sun serves as an Associate Editor for IEEE Transactions on Intelligent Vehicles, IEEE Robotics and Automation Letters, IEEE International Conference on Robotics and Automation, and IEEE/RSJ International Conference on Intelligent Robots and Systems.

Authorized licensed use limited to: CITY UNIV OF HONG KONG. Downloaded on November 30,2024 at 10:16:53 UTC from IEEE Xplore. Restrictions apply.