NLE-DM: Natural-Language Explanations for Decision Making of Autonomous Driving Based on Semantic Scene Understanding

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

***Abstract*— In recent years, the advancement of deep-learning technologies has greatly promoted the research progress of autonomous driving. However, deep neural network is like a black box. Given a specific input, it is difficult to explain the output of the network. Without explainable results, it would be unsafe to deploy deep networks in unseen environments or environments with potential unexpected situations. Especially for decision-making networks, inappropriate outputs could lead to severe traffic accidents. To provide a solution to this problem, we propose a deep neural network that jointly predicts the decision-making actions and corresponding natural-language explanations based on semantic scene understanding. Two types of explanations, the reasons of driving actions and the surrounding environment descriptions of the ego-vehicle, are designed. Both the reasons and descriptions are in the form of natural language. The decision-making actions could be explained by the corresponding reasons or the environment descriptions. We also release a large-scale dataset with hand-labelled ground truth including driving actions and environment descriptions. The superiority of our network over other methods is demonstrated on both our dataset and a public dataset.**

***Index Terms*— Autonomous driving, decision making, explainable artificial intelligence, semantic scene understanding.**

I. Introduction

AUTONOMOUS driving can reduce traffic accidents and improve driving safety, it has attracted great attention in the robotics and computer vision research communities in recent years. According to the survey of the American National Highway Traffic Safety Administration (NHTSA), around 94% of road accidents are caused by human factors [1], such as attention distraction, violation of traffic rules, etc. Since autonomous driving can eliminate human factors, it can greatly improve driving safety. Although significant research progress has been made in the past decade, autonomous driving is still not mature. The solutions based on traditional techniques still have not made great progress. Deep learning-based artificial intelligence has achieved great success in recent years. It has been widely applied in various research fields. With deep learning, autonomous driving technologies have been greatly advanced. Many effective deep neural networks for various autonomous driving applications, such as object detection [2], semantic scene understanding [3], [4], localization [5], motion planning [6], [7], trajectory prediction [8], [9], vehicle control [10], [11], and decision making [12], have been proposed.

Decision making is a process that selects one action from a set of discrete control actions (e.g., going straight, or turning left/right) based on the status of the ego-vehicle and the surrounding environment information [13]. It is an important component in autonomous driving. In recent years, many methods for decision making have been proposed. We can generally divide the existing methods into classical methods and deep learning-based methods. For the classical methods, there are rule-based methods, optimization-based methods and probabilistic methods, etc. Real traffic environments are often complex and dynamic. Compared to classical methods, deep learning-based methods could produce better performance in real environments [13], [14].

Despite the success of deep learning-based methods, their outputs are generally not explainable. The major reason is that deep neural network is like a black box. It is difficult to understand why they produce an output given a specific input. The lack of explainability has hindered them from being deployed in real-world environments, because real-world environments are dynamic, complex and unpredictable. The decision-making actions could not be anticipated given as input random sensory data captured in real-world environments, especially in unseen environments or when there are unexpected disturbances in the scene. Thus, it is really unsafe to believe the output of the deep networks without explainability.

To provide a solution to this problem, we propose a deep neural network that jointly predicts decision-making actions and natural-language explanations based on semantic scene understanding. Two types of explanations, the reasons of driving actions as well as the descriptions of surrounding environment of ego-vehicle, are proposed to explain the decision-making actions. To train and evaluate the network that jointly predicts decision-making actions and environment

Manuscript received 3 July 2022; revised 30 October 2022 and 4 March 2023; accepted 10 April 2023. Date of publication 5 June 2023; date of current version 30 August 2023. This work was supported in part by the Zhejiang Lab under Grant 2021NL0AB01, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, in part by the National Natural Science Foundation of China under Grant 62003286, and in part by The Hong Kong Polytechnic University Start-up Fund under Grant P0034801. The Associate Editor for this article was J. Alvarez. *(Corresponding author: Yuxiang Sun.)*

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

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

Digital Object Identifier 10.1109/TITS.2023.3273547

descriptions, a large-scale dataset that contains 10, 000 images from the BDD-OIA [15] is annotated with 4 driving actions and 6 descriptions. Furthermore, to verify the prediction performance and generalization capability of the proposed network, 1, 500 driving frames are selected from the nuScenes dataset [16] and labelled with 4 driving actions and corresponding natural-language explanations (including 21 reasons and 6 descriptions). The superiority of our proposed network over the other methods is demonstrated on the publicly available dataset [15] and our datasets. The comparative experimental results demonstrate that both the prediction performance of the decision-making actions and explainability of the network is notably improved. The main contributions of this work are summarized as follows:

- A novel explainable decision-making network based on semantic scene understanding for autonomous driving is proposed. In this network, both the natural-language reasons of driving actions and the surrounding environment descriptions of ego-vehicle are applied to explain the decision-making actions.
- A large-scale dataset that contains 10, 000 images with hand-labelled driving actions and descriptions of driving environments is released. Moreover, our code is open-sourced.[1]
- The superiority of our proposed network over other networks is demonstrated on both the publicly available dataset and our released datasets.

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 Artificial Intelligence

Explainable artificial intelligence (XAI) aims to enable humans to understand and appropriately trust AI algorithms [17], [18]. So far, a number of XAI techniques have been proposed and applied in different machine learning models on different tasks, including transparent model [19], [20], local explanation [21], [22], explanation by simplification [23], feature relevance explanation [24], visual explanation [25], [26], [27], architecture modification [28], [29], [30], [31], [32], etc.

The transparent model refers to the model that is explainable by itself. According to the degrees of explainability, it can be generally divided into three categories: simulatable model, decomposable model and algorithmically transparent model [33]. For local explanation, segmentation of the solution space is applied so that the explanation for the less complex solution subspaces can be obtained [18]. For explanation by simplification, the local interpretable model-agnostic explanations (LIME) and its variations [34] are the widely-used techniques. The main idea of LIME is to build the local linear models around the predictions. For feature relevance explanation, it gives the explanation of inner function of a model by assigning and calculating the relevance score of input features based on the importance of each feature in predicting a target variable [18]. For visual explanation, it is usually applied along with other techniques to visualize the behavior of model so that humans could have a better understanding of the model. For architecture modification, many different techniques can be performed to modify the architecture of the network so that the explainable network is obtained. These techniques includes layer modification [28], model combination [29], attention networks [30], [31], loss modification [32], etc. Take the work [30] as an example. The network architecture is modified by adding a global average pooling layer between the last convolutional layer and the fully-connected layer. With this architecture modification, the attention map that highlights the image regions that are particularly related to a specific object class is proposed. In our work, to give explanations to the actions, the decision-making network of autonomous driving is modified by adding the explanation module that predicts the natural-language explanations of the actions. The loss of the network is also modified into the multi-task loss so that the proposed network is able to jointly predict the driving actions and explanations. So, our work can be classified as an architecture modification method.

B. Explainable Autonomous Driving Systems

Considering the fact that the explainability is vital for autonomous driving, explainable autonomous driving systems (EADS) have attracted great interest among the research community. To improve the performance of EADS, many efforts [15], [35], [36], [37], [38], [39], [40], [41], [42] have been made in this area. Hofmarcher et al. [35] proposed an architecture that delivers real-time viable segmentation performance, which can be used as an input for an interpretable autonomous driving. Cultrera et al. [36] proposed to train an imitation learning based agent equipped with an attention model. Li et al. [37] added XAI technology in system to explain and assist the estimation results in the risk assessment phase. Shen et al. [38] focused on when an explanation is needed and how the content of explanation changes with context in autonomous driving. Atakishiyev et al. [39] presented a framework that integrates the autonomous driving, XAI architecture, and regulatory compliance to address this issue.

C. Datasets for Autonomous Driving

Autonomous driving largely depends on real-world datasets to develop, test and verify algorithms before deployment on public roads. So far, there are a number of autonomous driving datasets [15], [16], [43], [44], [45], [46], [47] that contain vision-based information or the information from multiple sensors, including GPS, radar, LiDAR, or IMU information. KITTI [43] is focused on the tasks of stereo, optical flow, visual odometry and 3D object detection. CityScapes [44] is aimed to assess the performance for semantic urban scene understanding tasks with high-quality annotations of 5k frames and 20k weakly annotated frames. Apolloscape [45] contains 144k frames from 4 regions in China under different times of

[1]Our code and dataset: https://github.com/lab-sun/NLE-DM

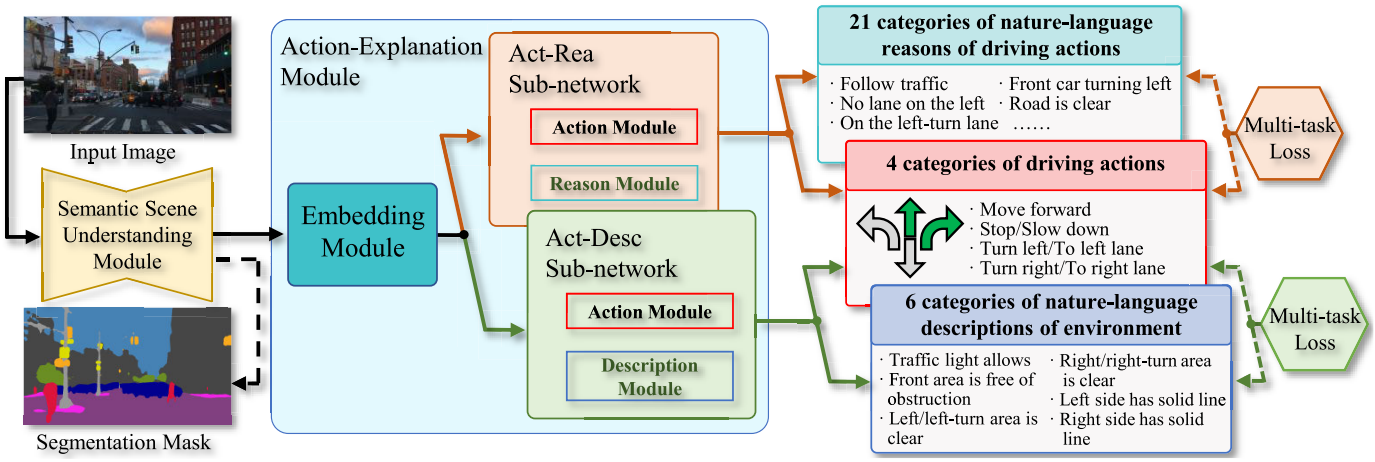

Fig. 1. The architecture of our proposed explainable network. We use the semantic scene understanding module for visual feature extraction. The action-explanation module takes as input the feature maps from the semantic scene understanding module, and produces the decision-making actions and the corresponding natural-language explanations. The figure is best viewed in color.

day and weather conditions. nuScenes [16] has 1.4M camera images, 390k LIDAR sweeps, 1.4M RADAR sweeps and 1.4M object bounding boxes in 1000 scenes. BDD100K [46] is one of the largest driving video dataset with 100K videos and multitasks including object detection, semantic segmentation, lane detection, etc. BDD-OIA [15] is a subset selected from BDD100K, which contains at least 5 pedestrians or bicycle riders and more than 5 vehicles. In the BDD-OIA dataset, ground truth for 4 actions and the corresponding 21 explanations are annotated.

D. Difference With Existing Work

The closest work to ours is the explainable object-induced action method proposed by Xu et al. [15]. The authors proposed a paradigm to focus on the action-inducing objects by combining action-inducing object reasoning and global scene reasoning. In their work, a multi-task network (we call the OIA network) along with a dataset (i.e., the BDD-OIA) was proposed to predict the actions and explanations. Unlike the OIA network [15], we employ semantic segmentation to understand the surrounding traffic environment, while the OIA network uses object detection. Compared with the OIA network, our network could capture more detailed information of the environment, which leads to better prediction performance of the explanation. Furthermore, we introduce the natural-language descriptions of surrounding environments to explain the driving actions, and release a new large-scale dataset with hand-labelled driving actions as well as the corresponding natural-language environment descriptions. Compared with the explanations in the OIA method, our natural-language descriptions are more concise and straightforward, which could produce better explainability for the driving actions.

III. The Proposed Network

In this section, we present the architecture of the proposed network and the training details of our network.

A. The Network Architecture

As illustrated in Fig. 1, our network could be mainly divided into two components: the semantic scene understanding module and the action-explanation module. Our network takes as input a front-view image and predicts the driving actions and the corresponding explanations.

Let $X_i \in \mathbb{R}^{h\times w\times c}$ denote the i-th image in a frame set $\mathbf{X} = \{X_1, \ldots, X_i, \ldots, X_N\}$, where N is the number of frames, h, w and c respectively denote the height, width and number of channels for an image. For the driving actions, 4 categories of actions are adopted: "move forward", "stop/slow down", "turn left/change to left lane", and "turn right/change to right lane". For the explanations, two types of natural-language explanations, that is, the reasons of driving actions and the surrounding environment descriptions, are proposed to explain the actions. Specifically, we adopt 21 categories of reasons and 6 categories of environment descriptions in this work. For the natural-language reasons, they have been used in the work [15] and all the 21 categories of the reasons are displayed in Tab. II. For the environment descriptions, we adopt "traffic light allows", "front area is free of obstruction", "left/left-turn area is clear", "right/right-turn area is clear", "left side has solid line" and "right side has solid line". The 6 environment descriptions are displayed in Tab. V and the details of the environment descriptions are discussed in Section IV. The whole process of our NLE-DM network is described as follows:

$$\mathbf{X} \rightarrow (A, R) \in \{0,1\}^4 \times \{0,1\}^{21},$$
$$\text{or}$$
$$\mathbf{X} \rightarrow (A, D) \in \{0,1\}^4 \times \{0,1\}^6, \quad (1)$$

where A denotes the driving actions, R and D denote the reasons of driving actions and the environment descriptions, respectively.

The semantic scene understanding (S-S) module is based on a semantic segmentation network, DeepLabv3 [48]. We refer readers to [48] for more details about DeepLabv3. The function of the S-S module is to extract the semantic feature

map $M_i \in \mathbb{R}^{h\times w\times n}$ from the image X_i, where h and w denote height and width of the feature map, which are the same as the image X_i, n denotes the number of channels for semantic feature map, which is also the number of classes for the semantic segmentation. The process of the S-S module is described as follows:

$$\text{S-S Module: } X_i \rightarrow M_i \in \mathbb{R}^{h\times w\times n}, 1 \leq i \leq N, \quad (2)$$

where N is the number of frames in the frame set $\mathbf{X}$.

The action-explanation (A-E) module consists of three parts: the embedding module, the Act-Rea sub-network and the Act-Desc sub-network. The feature maps from the S-S module are first fed into the embedding module to reduce the dimensionality and resolution. After the processing, the shape of the embedding feature map is $64 \times 18 \times 32$, where 64 is the number of channels, and 18×32 is the resolution. Note that in this work, the input resolution is 720×1280 and the number of channels is 3 (i.e., RGB image). Then, the embedding feature map is flattened and fed into the Act-Rea sub-network and the Act-Desc sub-network to predict the decision-making actions and the corresponding natural-language explanations. The process of embedding module is described as $M_i \rightarrow V_i$, where V_i is the flatten vector.

As aforementioned, the reasons of driving actions and the environment descriptions are both proposed to explain the decision-making actions. So, we design the Act-Rea sub-network and the Act-Desc sub-network in the A-E module to output the driving actions with different natural-language explanations. The Act-Rea sub-network contains the action and reason modules so that the A-E module outputs the decision-making actions along with the natural-language reasons of driving actions. The process for the Act-Rea sub-network is described as follows:

$$\text{Act-Rea: } V_i \rightarrow (A, R) \in \{0,1\}^4 \times \{0,1\}^{21}, 1 \leq i \leq N, \quad (3)$$

where N is the number of frames in the frame set $\mathbf{X}$. The Act-Desc sub-network contains the action and description modules so that the A-E module outputs the decision-making actions along with the natural-language descriptions of the surrounding environment of the ego-vehicle. The process for the Act-Desc sub-network is described as follows:

$$\text{Act-Desc: } V_i \rightarrow (A, D) \in \{0,1\}^4 \times \{0,1\}^6, 1 \leq i \leq N, \quad (4)$$

where N is the number of frames in the frame set $\mathbf{X}$.

B. Training Details

We first pre-train the S-S module using the BDD10K dataset, which is a part of the BDD100K dataset [46]. The S-S module enables our network to be equipped with the capability for pixel-wise semantic scene understanding. Then, we train our whole network by loading the pre-trained weight. For the Act-Rea sub-network, the training and testing is based on the BDD-OIA dataset [15]. For the Act-Desc sub-network, we proposed a new dataset, the *BDD Actions and Descriptions* (BDD-AD) dataset. The images of BDD-AD are selected from the BDD-OIA and annotated with driving actions and natural-language descriptions of the surrounding environment of the ego-vehicle. The details of the BDD-AD dataset are presented in Section IV. To further verify the prediction performance and generalization capacity of our proposed network, both the Act-Rea and Act-Desc are tested on 1,500 images that are selected from the nuScenes dataset [16] and labelled with driving actions and corresponding natural-language explanations (including reasons and descriptions).

We adopt the stochastic gradient descent (SGD) optimizer with the initial learning rate of 0.001, momentum of 0.9 and weight decay of 1×10^{-4}. It is worth noting that the images between BDD10K and BDD-OIA are not totally overlapped with each other. So, to enable our network to adapt to new scenes, even though the network is pre-trained on BDD10K, the weights of the S-S module are not fixed during training with BDD-OIA or BDD-AD.

Our network is trained with a multi-task loss function, which is calculated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{act} + \lambda \mathcal{L}_{rea}, \quad (5)$$

$$\mathcal{L}_{total} = \mathcal{L}_{act} + \lambda \mathcal{L}_{desc}, \quad (6)$$

where $\mathcal{L}_{total}$ is the total loss, $\mathcal{L}_{act}$, $\mathcal{L}_{rea}$ and $\mathcal{L}_{desc}$ are the binary cross entropy losses for action prediction, reason prediction and description prediction, respectively. For the Act-Rea sub-network, loss function (5) is applied. For the Act-Desc sub-network, loss function (6) is applied. λ is a weight parameter that determines the relative importance between decision-making actions and the corresponding natural-language explanations.

For the Act-Rea sub-network, we adopt the 4 categories of decision-making actions and corresponding 21 categories of natural-language reasons that are used in the work [15]. It is worth noting that in the work [15], the 21 categories of natural-language reasons are referred as "explanations". In such case, the two losses $\mathcal{L}_{act}$ and $\mathcal{L}_{rea}$ are calculated as: $\mathcal{L}_{act} = \sum_{i=1}^{4} \mathcal{L}[\hat{A}_i, A_i]$ and $\mathcal{L}_{rea} = \sum_{j=1}^{21} \mathcal{L}[\hat{R}_j, R_j]$, where $\hat{A}_i$ and A_i are the prediction and ground truth for decision-making actions, respectively. $\hat{R}_j$ and R_j are the prediction and ground truth for the reasons, respectively.

For the Act-Desc sub-network, the 4 categories of decision-making actions are still adopted but with 6 categories of natural-language descriptions. Therefore, besides the $\mathcal{L}_{act}$, the $\mathcal{L}_{desc}$ is calculated as: $\mathcal{L}_{desc} = \sum_{k=1}^{6} \mathcal{L}[\hat{D}_k, D_k]$, where $\hat{D}_k$ and D_k are the prediction and ground truth for the descriptions, respectively.

IV. Experimental Results and Discussions

A. Evaluation Metrics

The standard F1 score metric is employed to quantitatively evaluate the prediction performance of decision-making actions, the prediction performance of the reasons of driving actions, and the prediction performance of the descriptions of the surrounding environment. Two types of F1 score, $F1_{oval}$ and $F1_m$, are used. The $F1_{oval}$ refers to the overall F1 score,

TABLE I

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

Methods	F	S	L	R	$F1_m^{act}$	$F1_{oval}^{act}$	$F1_m^{rea}$	$F1_{oval}^{rea}$
Act-Rea ($\lambda = 1.0$)	0.827	0.760	0.651	0.653	**0.723**	*0.733*	0.312	0.517
Act-Rea ($\lambda = 2.0$)	0.813	0.768	0.649	0.643	*0.718*	0.728	**0.350**	**0.546**
OIA [15]	0.829	0.781	0.630	0.634	*0.718*	**0.734**	0.208	0.422
Local selector [49]	0.810	0.762	0.600	0.624	0.699	0.711	0.196	0.406
C-SENN [41]	0.772	0.744	0.469	0.486	0.618	–	–	–
CBM [50]	0.795	0.732	0.483	0.431	0.610	0.661	0.292	0.412
CBM-AUC [42]	0.803	0.751	0.551	0.525	0.658	0.704	*0.342*	*0.522*

which is calculated as:

$$F1_{oval}^{act} = \frac{1}{N}\sum_{i=1}^{N} F1(\hat{A}_i, A_i), \tag{7}$$

$$F1_{oval}^{rea} = \frac{1}{M}\sum_{j=1}^{M} F1(\hat{R}_j, R_j), \tag{8}$$

$$F1_{oval}^{desc} = \frac{1}{Q}\sum_{k=1}^{Q} F1(\hat{D}_k, D_k), \tag{9}$$

where $F1_{oval}^{act}$, $F1_{oval}^{rea}$ and $F1_{oval}^{desc}$ are the $F1_{oval}$ scores for the action predictions, the reason predictions and the description predictions, respectively. N, M and Q are the numbers of the action predictions, the reason predictions and the description predictions, respectively. Considering the fact that the BDD-OIA and BDD-AD datasets are imbalanced, we also calculate the mean F1 score, $F1_m$, as follows:

$$F1_m^{act} = \frac{1}{4}(F1_F + F1_S + F1_L + F1_R), \tag{10}$$

$$F1_m^{rea} = \frac{1}{21}\sum_{j=1}^{21} F1_j^{rea}, \tag{11}$$

$$F1_m^{desc} = \frac{1}{6}\sum_{k=1}^{6} F1_k^{desc}, \tag{12}$$

where $F1_m^{act}$, $F1_m^{rea}$ and $F1_m^{desc}$ are the $F1_m$ scores for the action predictions, the reason predictions and the description predictions, respectively. $F1_F$, $F1_S$, $F1_L$ and $F1_R$ are F1 scores for the predictions of "move forward", "stop/slow down", "turn left/change to left lane" and "turn right/change to left lane", respectively. $F1_j^{rea}$ and $F1_k^{desc}$ are the F1 score for the predictions of each reason category and description category, respectively.

B. Jointly Predicting Actions and Reasons

In this section, we discuss the experimental results for the Act-Rea sub-network to jointly predict the decision-making actions and the corresponding natural-language reasons. Tab. I shows the quantitatively comparative results between our Act-Rea sub-network and other networks [15], [41], [42], [49], [50]. As aforementioned, the λ is the weighting parameter in the loss function (5) to determine the relative importance between driving actions and corresponding reasons. The effectiveness of λ on the prediction performance is discussed in detail in the ablation study. The OIA network [15] combines the object reasoning with global scene reasoning to focus on the action-inducing object. The network of local selector is proposed by Wang et al. [49] and modified by Xu et al. [15], and it is able to predict the action and explanation. The local selector could be regarded as the OIA network that contains only the local object reasoning. The contrastive self-explaining neural network (C-SENN) [41] combines contrastive learning with concept learning to improve the explainability and the accuracy of predictions of driving actions. The concept bottleneck model (CBM) [50] is proposed by Kon et al. [50] and modified by Sawada et al. [42] to jointly predict action and corresponding reasons. The concept bottleneck model with additional unsupervised concepts (CBM-AUC) [42] is based on CBM and integrates supervised concepts with unsupervised concepts to improve prediction performance.

As shown in Tab. I, for the action prediction, our Act-Rea ($\lambda = 1.0$ & 2.0) and OIA have similar prediction performance,

TABLE II

THE PREDICTION PERFORMANCE OF THE NATURAL-LANGUAGE REASONS. WITH THE BDD-OIA DATASET, TO ALLEVIATE THE IMBALANCE, "TURN LEFT/RIGHT" IS MERGED WITH "CAN'T CHANGE TO LEFT/RIGHT LANE". HERE, FOR THE CONVENIENCE OF ILLUSTRATION, "TURN LEFT/RIGHT" AND "CAN'T CHANGE TO LEFT/RIGHT LANE" ARE LISTED IN DIFFERENT ROWS

Action Category	Reason Category	F1 Score
Move forward	follow traffic	0.645
	the road is clear	0.447
	the traffic light is green	0.528
Stop/Slow down	obstacle: car	0.599
	obstacle: person/pedestrian	0.440
	obstacle: rider	0.000
	obstacle: others	0.000
	the traffic light	0.768
	the traffic sign	0.000
Turn left	front car turning left	0.000
	on the left-turn lane	0.000
	traffic light allows	0.000
Turn right	front car turning right	0.000
	on the right-turn lane	0.053
	traffic light allows	0.000
Can't change to left lane	obstacles on the left lane	0.585
	no lane on the left	0.472
	solid line on the left	0.474
Can't change to right lane	obstacles on the right lane	0.624
	no lane on the right	0.474
	solid line on the right	0.442

Fig. 2. Sample comparative results of action and reason predictions for the OIA network [15] and our network. The label "GT" denotes the ground truth for decision-making actions, and the label "Pre" denotes the prediction for decision-making actions. For the reasons, green denotes true positive, red denotes false positive, grey denotes false negative. To ensure the fairness of comparison, all the four figures are chosen exactly the same with the Fig. 4 from [15]. The figure is best viewed in color.

TABLE III
THE PREDICTED IOU (%) FOR EACH CLASS ON THE BDD10K DATASET

Class Name	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain
IoU (%)	94.2	64.2	84.7	41.6	52.1	36.9	46.4	47.2	85.7	48.4
Class Name	sky	person	rider	car	truck	bus	train	motocycle	bicycle	
IoU (%)	94.5	59.8	13.4	89.5	57.1	78.5	0.00	36.3	36.2	

which are better than the other networks. For the reason prediction, the performance of our Act-Rea ($\lambda = 1.0$ & 2.0) and CBM-AUC are at the same level and are better than the other networks. Therefore, these comparative results quantitatively demonstrate the superiority of our proposed Act-Rea sub-network in terms of the prediction performance of both the decision-making actions and the corresponding reasons.

The superiority of our network is also validated by the qualitative results as shown in Fig. 2. In order to ensure the fairness of the comparison, we choose exactly the same examples from the paper of OIA [15]. As shown in Fig. 2, the decision-making action predictions of Act-Rea ($\lambda = 1.0$) sub-network and OIA network are the same, but with different reason prediction accuracy. For the OIA network, the ratio of true positive for the reason prediction of four examples is 100%, 33.3%, 50%, 66.7%, respectively. For our Act-Rea sub-network, the ratios are 100%, 50%, 75%, 100%.

We conjecture the reasons why the prediction performance of our network is better than OIA as follows:

- Unlike our network that uses the atrous spatial pyramid pooling (ASPP) [48] to capture multi-scale features, the OIA network only involves global and local features. Given the fact that the driving environment is complex with various sizes of objects, the lack of multi-scale environment perception may cause the network to misidentify multi-scale objects and eventually lead to incorrect reason predictions.
- In the OIA network, the object detection (i.e., Faster R-CNN) is employed to obtain the feature maps and capture the action-induced objects. By contrast, our network uses semantic segmentation (i.e., DeepLabv3) instead of object detection, which could capture more detailed information from the environments, because semantic segmentation predicts object class labels at the fine pixel-wise level while object detection provides the labels only at the coarse bounding-box level.

The prediction performance for the natural-language reason is also discussed in this section. Tab. II shows the prediction performance of each category of the reasons of

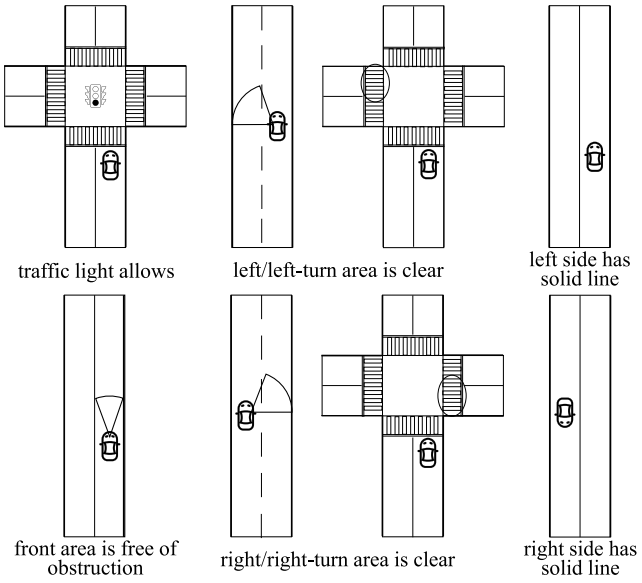

Fig. 3. The schematic diagram for surrounding environment descriptions of the ego-vehicle.

TABLE IV
COMPARATIVE RESULTS ON BDD-OIA AND NU-AR FOR THE PREDICTION PERFORMANCE OF THE ACT-REA SUB-NETWORK

Test Set	$F1_{m}^{act}$	$F1_{oval}^{act}$	$F1_{m}^{rea}$	$F1_{oval}^{rea}$
BDD-OIA	0.723	0.733	0.312	0.517
nu-AR	0.688	0.722	0.308	0.499

the Act-Rea ($\lambda = 1.0$). Unlike the prediction performance of the decision-making action, the reason prediction results are biased. For some reason categories, their F1 scores are zero, which indicates that the network is unable to predict the natural-language reasons. The reasons for unsatisfactory natural-language reason predictions are conjectured as follows:

- The network is pre-trained on the BDD10K dataset for semantic segmentation. Naturally, the poor segmentation performance for some object classes from the semantic segmentation could result in unsatisfactory reason prediction. For example, the intersection over union (IoU) for the rider class is 13.4% (see details in Tab. III), which may cause the A-E module in the network to misidentify the "rider", not to mention using "obstacle: rider" to explain the driving action of "Stop/Slow down". A similar reason for the unsatisfactory prediction performance of the reason of "obstacle: others", as the IoU of other obstacles, including train, motorcycle and bicycle, are 0.0%, 36.3% and 36.2%, respectively.
- Some natural-language reasons are abstract and ambiguous, which may also cause incorrect predictions. Taking the natural-language reasons of "front car turning left/right" as an example, even though the IoU for the car class is satisfactory (89.5%), the network still fails the reason prediction, because the network could not understand whether the front car is turning left/right or not.

To further test the prediction performance and the generalization capability of the Act-Rea sub-network, we select

TABLE V
THE CATEGORIES OF THE ACTIONS AND DESCRIPTIONS IN OUR PROPOSED BDD-AD DATASET. THE RATIO REFERS TO THE PERCENTAGE OF EACH CATEGORY IN THE DATASET

Annotation	Category	Ratio
Action	Move forward	73.68%
	Stop/slow down	24.78%
	Turn left/change to left lane	39.42%
	Turn right/change to right lane	44.34%
Description	Traffic light allows	73.02%
	Front area is free of obstruction	82.37%
	Left/left-turn area is clear	65.00%
	Right/right-turn area is clear	59.10%
	Left side has solid line	28.83%
	Right side has solid line	18.15%

1,500 images from the nuScenes dataset [16] and label each image with the 4 driving actions and 21 reasons. This dataset is named as the *nuScenes Actions and Reasons* (nu-AR) dataset. In nu-AR, both the driving action categories and reason categories are the same as the BDD-OIA dataset [15]. Then, we load the weight that is trained on BDD-OIA and obtain the prediction performance of the Act-Rea sub-network that is tested on nu-AR. Tab. IV shows the comparative results of prediction performance for the Act-Rea ($\lambda = 1.0$) that is tested on BDD-OIA and nu-AR, respectively. As shown in Tab. IV, both the action and description prediction performance of Act-Rea tested on BDD-OIA is slightly better than those of the Act-Rea tested on nu-AR. For the $F1_{m}^{act}$ and $F1_{oval}^{rea}$, the testing results of BDD-OIA are about 5% higher than the testing results of nu-AR. For the $F1_{oval}^{act}$ and $F1_{m}^{rea}$, the testing results of BDD-OIA are about 1% higher than the testing results of nu-AR. These comparative results verify the prediction performance and generalization capability of our proposed Act-Rea sub-network.

C. Jointly Predicting Actions and Descriptions

To further improve the explainability of decision-making actions, here we propose to apply the natural-language descriptions of surrounding environment of the ego-vehicle to explain the decision-making actions. As shown in Fig. 3, to give a comprehensive description of surrounding environment, natural-language descriptions of "traffic light allows", "front area free of obstruction", "left/left-turn area is clear", "right/right-turn area is clear", "left side has solid line" and "right side has solid line" are chosen. For the description category of "left/left-turn area is clear", it contains the "left area of ego-vehicle is clear" and "left-turn area of crossroads is clear". For the description category of "right/right-turn area is clear", it contains the "right area of ego-vehicle is clear" and "right-turn area of crossroads is clear". Considering the fact that the relative ratios of "left/right-turn area of crossroads is clear" are low, the description categories of "left/right area of ego-vehicle is clear" and "left/right-turn area of crossroads is clear" are merged into "left/left-turn (right/right-turn) area is clear" to avoid biased distribution. It is clear to see that compared with the natural-language reasons of driving actions,

TABLE VI
COMPARATIVE RESULTS OF THE PREDICTION PERFORMANCE FOR THE ACT-DESC SUB-NETWORK AND ACT-REA SUB-NETWORK

Networks	$F1_{m}^{act}$	$F1_{oval}^{act}$	$F1_{m}^{desc}$	$F1_{oval}^{desc}$	$F1_{m}^{rea}$	$F1_{oval}^{rea}$
Act-Desc	0.876	0.877	0.907	0.880	–	–
Act-Rea	0.723	0.733	–	–	0.312	0.517

TABLE VII
COMPARATIVE RESULTS ON BDD-AD AND NU-AD FOR THE PREDICTION PERFORMANCE OF THE ACT-DESC SUB-NETWORK

Test Set	$F1_{m}^{act}$	$F1_{oval}^{act}$	$F1_{m}^{desc}$	$F1_{oval}^{desc}$
BDD-AD	0.876	0.877	0.907	0.880
nu-AD	0.760	0.836	0.882	0.879

the surrounding environment descriptions of the ego-vehicle are much more straightforward and objective.

In order to use the surrounding environment descriptions to explain the decision-making actions, the Act-Desc sub-network jointly predicts the decision-making actions and natural-language environment descriptions. Then, the descriptions could be used to explain the decision-making actions. For example, if the natural-language descriptions of the surrounding environment are "traffic light allows" and "front area is free of obstruction", then the action of "move forward" could be explained.

To train the Act-Desc sub-network to jointly produce decision-making actions and the description of the ego-vehicle's surrounding environment, a large-scale dataset with hand-labelled driving actions and natural-language descriptions of the ego-vehicle's surrounding environment is built by us. We refer this dataset as *BDD Actions and Descriptions* (BDD-AD) dataset, because $10,000$ images from the BDD-OIA dataset [15] are selected from various weather conditions and different times of the day. Each image in BDD-AD is manually annotated with 4 driving actions ("move forward", "stop/slow down", "turn left/change to left lane", "turn right/change to right lane") and 6 natural-language descriptions of the ego-vehicle's surrounding environment. In addition, each image in BDD-AD contains at least 5 pedestrians or bicycle riders and more than 5 vehicles. Therefore, considering the complex driving scenes, multiple driving actions and descriptions are annotated for each image. Tab. V summarizes the number of each category of actions and descriptions. We use binary vectors to represent the driving actions and natural-language descriptions. For example, if the driving actions of the ego-vehicle are "move forward" and "change to right lane", and the surrounding environment descriptions are: "traffic light allows", "front area is free of obstruction", "left/left-turn area is clear", "right/right-turn area is clear", "left side has solid line" and "right side has no solid line", the annotations for the actions and descriptions are $[1, 0, 0, 1]^T$ and $[1, 1, 1, 1, 1, 0]^T$, respectively.

Tab. VI shows the comparative results between the Act-Rea ($\lambda = 1.0$) and the Act-Desc ($\lambda = 1.0$). For the Act-Rea sub-network, it jointly predicts decision-making actions and the corresponding natural-language reasons. For the Act-Desc sub-network, it jointly produces decision-making actions and the corresponding natural-language descriptions of ego-vehicle's surrounding environment. Since the images of BDD-AD dataset are selected from the BDD-OIA dataset, these two datasets have the same level of scene complexity and traffic conditions. Therefore, even though the dataset for Act-Rea sub-network and Act-Desc sub-network are different (Act-Rea sub-network is based on BDD-OIA dataset, while Act-Desc sub-network is based on BDD-AD dataset), the prediction performance between these two sub-networks are still comparable. As shown in Tab. VI, for the prediction performance of the decision-making actions, both the $F1_{m}^{act}$ and $F1_{oval}^{act}$ of the Act-Desc sub-network is about 20% higher than the Act-Rea sub-network. For the prediction performance of the natural-language explanations, the $F1_{m}^{desc}$ of Act-Desc sub-network is about 200% higher than the $F1_{m}^{rea}$ of Act-Rea sub-network, and the $F1_{oval}^{desc}$ is about 70% higher than $F1_{oval}^{rea}$. Therefore, compared with Act-Rea sub-network, the Act-Desc sub-network has better prediction performance both in decision-making actions and natural-language explanations.

The possible reasons why the prediction performance of the decision-making actions and natural-language explanations of the Act-Desc sub-network are better than the Act-Rea sub-network are discussed as follows:

- Compared with some natural-language reasons (such as "follow traffic", "front car turning left/right", "on the left/right turn lane", etc), all the 6 natural-language descriptions of ego-vehicle's surrounding environment are more straightforward and precise, which could lead to better prediction performance of natural-language explanation for the Act-Desc sub-network.
- It can be discovered that the existence of natural-language explanations could improve the prediction performance of decision-making actions (see the details in the following ablation study). So, the better prediction performance of natural-language explanation of Act-Desc sub-network also leads to more accurate prediction of decision-making actions.

Fig. 4 shows sample qualitative results for the Act-Desc sub-network to jointly predict the decision-making actions and the corresponding descriptions of the surrounding environment. Different weather conditions and different times of the day are chosen to demonstrate the generalization capability of our proposed network. As shown in Fig. 4, all the decision-making action predictions and most description predictions of the Act-Desc sub-network are the same as ground truth, which demonstrates the satisfactory performance of the Act-Desc sub-network.

To further test the prediction performance and the generalization capability of the Act-Desc sub-network, we label the selected 1.5k nuScenes images with 4 driving actions and 6 natural-language descriptions. This dataset is named as the *nuScenes Actions and Descriptions* (nu-AD) dataset. In nu-AD, both the driving action categories and description categories are the same as the BDD-AD dataset. Then, we load the weight that is trained on BDD-AD and obtain the prediction performance of the Act-Desc sub-network that is tested on nu-AD. Tab. VII shows the comparative results

Fig. 4. The sample prediction results of the decision-making actions and the surrounding environment descriptions of the ego-vehicle. The label "GT" denotes the ground truth for action, and the label "Pre" denotes the action prediction. For the descriptions, green explanation denotes correct predictions, red explanation denotes wrong predictions. The figure is best view in color.

of prediction performance for the Act-Desc ($\lambda = 1.0$) that is tested on BDD-AD and nu-AD, respectively. The action prediction performance of the Act-Desc tested on BDD-AD is about 10% better than the Act-Desc tested on nu-AD. The description prediction performance of the Act-Desc tested on BDD-AD is slightly better than the Act-Desc tested on nu-AD. For the $F1_{m}^{desc}$, the testing result of BDD-AD is about 3% higher than the testing result of nu-AD. For the $F1_{oval}^{desc}$, the testing results of BDD-AD and nu-AD are almost the same. These comparative results verify the prediction performance and generalization capability of our proposed Act-Desc sub-network.

D. Ablation Study

We first investigate the relationship between the decision-making actions and the corresponding natural-language explanations. Tab. VIII shows the prediction performance of the Act-Rea sub-network with different values of λ in the loss function (5). As aforementioned, the weighting parameter λ determines the relative importance between decision-making actions and corresponding reasons. The Act-Rea sub-network with $\lambda = 0.0$ means that the reason prediction is removed. On the contrary, the $\lambda = \infty$ refers to the Act-Rea sub-network without action prediction. As shown in Tab. VIII, for the Act-Rea with $\lambda = 0.0$

TABLE VIII

THE ABLATION STUDY RESULTS OF PREDICTION PERFORMANCE FOR ACT-REA SUB-NETWORKS WITH THE DIFFERENT RELATIVE IMPORTANCE OF ACTION AND REASON. THE RELATIVE IMPORTANCE IS DETERMINED BY λ ON THE LOSS FUNCTION (5). THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT AND ITALIC FONT

λ	F	S	L	R	$F1_m^{act}$	$F1_{oval}^{act}$	$F1_m^{rea}$	$F1_{oval}^{rea}$
0.0	0.808	0.710	0.609	0.631	0.690	0.697	–	–
0.5	0.815	0.769	0.646	0.644	*0.718*	0.725	0.302	0.506
1.0	0.827	0.760	0.651	0.653	**0.723**	**0.733**	0.312	0.517
2.0	0.813	0.768	0.649	0.643	*0.718*	*0.728*	*0.350*	*0.546*
∞	–	–	–	–	–	–	**0.372**	**0.568**

TABLE IX

THE ABLATION STUDY RESULTS OF THE ACT-DESC SUB-NETWORK. THE PREDICTION PERFORMANCE FOR THE ACT-DESC SUB-NETWORKS WITH THE DIFFERENT RELATIVE IMPORTANCE OF ACTION AND DESCRIPTION (TOP). THE PREDICTION PERFORMANCE FOR NETWORKS WITH DIFFERENT BACKBONES (BOTTOM). THE RELATIVE IMPORTANCE OF ACTION AND DESCRIPTION IS DETERMINED BY λ ON THE LOSS FUNCTION (6). THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN BOLD FONT AND ITALIC FONT

λ /Encoder	F	S	L	R	$F1_m^{act}$	$F1_{oval}^{act}$	$F1_m^{desc}$	$F1_{oval}^{desc}$
0.0	0.913	0.753	0.725	0.751	0.785	0.794	–	–
0.5	0.941	0.845	0.830	0.836	*0.863*	*0.865*	0.900	0.876
1.0	0.949	0.858	0.832	0.866	**0.876**	**0.877**	0.907	0.880
2.0	0.938	0.848	0.829	0.836	0.862	0.863	*0.909*	*0.889*
∞	–	–	–	–	–	–	**0.921**	**0.894**
ResNet50	0.949	0.858	0.832	0.866	0.876	0.877	*0.907*	0.880
ResNet101	0.955	0.859	0.836	0.870	**0.880**	**0.881**	**0.916**	**0.892**
MobileNetV3-Small	0.921	0.782	0.780	0.788	0.818	0.821	0.824	0.827
MobileNetV3-Large	0.950	0.854	0.840	0.870	*0.878*	*0.880*	0.903	*0.885*

(only action prediction), the prediction performance of the decision-making actions is worse than the Act-Rea with $\lambda = 1.0$ (both action and reason predictions), which indicates that the existence of the reasons could improve the prediction performance of the decision-making actions. For the Act-Rea with $\lambda = 0.5$, its action prediction performance is better than the Act-Rea with $\lambda = 0.0$ and worse than the Act-Rea with $\lambda = 1.0$, which again validates that the existence of reason could impose positive impacts on the action prediction. However, it is worth noting that the positive impacts of reason on action prediction are limited, because the action prediction performance of the Act-Rea with $\lambda = 2.0$ is worse than the Act-Rea with $\lambda = 1.0$.

For the Act-Rea with $\lambda = \infty$ (only reason prediction), its prediction performance of the natural-language reasons is better than the Act-Rea with $\lambda = 1.0$, indicating that the existence of action has no positive effect on the reason prediction. Furthermore, the reason prediction performance of the Act-Rea with $\lambda = 2.0$ is better than the Act-Rea with $\lambda = 1.0$ and worse than the Act-Rea with $\lambda = \infty$, which again validates that the existence of action could not improve the reason prediction.

A similar relationship between the decision-making actions and descriptions of the surrounding environment is also discovered for the Act-Desc sub-network (see the top rows of Tab. IX). For the Act-Desc with $\lambda = 0.0$ (only action prediction), the prediction performance of the decision-making actions is worse than the Act-Desc with $\lambda = 1.0$ (both action and description predictions), which indicates that the existence of the description could improve the prediction performance of the decision-making actions. For the Act-Desc with $\lambda = 0.5$, its action prediction performance is better than the Act-Desc with $\lambda = 0.0$ and worse than the Act-Desc with $\lambda = 1.0$, which again validates that the existence of description could impose positive impacts on the action prediction. For the Act-Desc with $\lambda = \infty$ (only description prediction), its prediction performance of description is better than the Act-Desc with $\lambda = 1.0$, indicating that the existence of action has no positive effect on the description prediction. The description prediction performance of the Act-Desc with $\lambda = 2.0$ is better than the Act-Desc with $\lambda = 1.0$ and worse than the Act-Desc with $\lambda = \infty$, which again validates that the existence of action could not improve the description prediction.

We think this is caused by the internal relationship between the decision-making actions and the corresponding explanations. Even though the action and explanation predictions are parallel from the view of the network architecture, there may exist some interactions or impacts between them. The decision-making action could be regarded as the result of the corresponding explanation. So, accurate explanation prediction could improve the action prediction, while the action predictions do not impose such impacts on the description predictions.

In this section, we also test different feature extraction backbones, including ResNet50 (baseline), ResNet101 [51], MobileNetV3-Small and MobileNetV3-Large [52], in the Act-Desc sub-network ($\lambda = 1.0$) to compare their prediction performance. As shown in the bottom rows of Tab IX, the Act-Desc sub-network with the ResNet101 backbone presents the best prediction performance in both the decision-making actions and surrounding environment descriptions. For the Act-Desc sub-network with the MobileNetV3-Large backbone, its prediction performance is at the same level as the Act-Desc sub-network with the ResNet50 backbone. Even though the prediction performance of the Act-Desc sub-network with the MobileNetV3-Small backbone is the worst

among these networks, it is still acceptable, indicating that the Act-Desc sub-network has a potential to be applied on resource-constrained mobile devices.

E. Limitations

Despite the superiority of our proposed network, there are still some limitations. Firstly, the proposed network uses only one image frame. However, human drivers usually use a sequence of visual information to make driving decisions. So, using a sequence instead of a single image frame may improve the prediction performance of decision-making action. Moreover, for both the Act-Rea sub-network and the Act-Desc sub-network, the predicted decision-making actions are selected from only 4 action categories. More categories of decision-making actions should be considered to enable our network to work in real environments.

V. Conclusion and Future Work

We proposed here an explainable network to explain the decision-making actions by jointly predicting decision-making actions and the corresponding natural-language explanations for autonomous driving. Two types of explanations, the reasons of driving actions as well as the descriptions of ego-vehicle's surrounding environment, are proposed. We also release a large-scale dataset with hand-labelled ground truth including 4 kinds of driving actions and 6 kinds of natural-language descriptions of surrounding environment of the ego-vehicle. The comparative experiments are performed and the superiority of our network over other methods is demonstrated on both our datasets and a public dataset. The relationship between the decision-making actions and the corresponding natural-language explanations has been discussed through ablation studies. In the future, we would like to further enhance the explainability of the proposed NLE-DM by revealing the internal working principles of the network, which may improve the explainability of the network.

References

[1] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Nat. Highway Traffic Saf. Admin. (NHTSA), Washington, DC, USA, Tech. Rep. DOT HS 812 115, 2015.

[2] 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.

[3] Z. Feng et al., "MAFNet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022.

[4] H. Wang et al., "SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 21405–21417, Nov. 2022.

[5] M. U. M. Bhutta, Y. Sun, D. Lau, and M. Liu, "Why-so-deep: Towards boosting previously trained models for visual place recognition," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1824–1831, Apr. 2022.

[6] 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.

[7] B. Li, Y. Ouyang, L. Li, and Y. Zhang, "Autonomous driving on curvy roads without reliance on Frenet frame: A Cartesian-based trajectory planning method," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15729–15741, Sep. 2022.

[8] Y. Sun, W. Zuo, and M. Liu, "See the future: A semantic segmentation network predicting ego-vehicle trajectory with a single monocular camera," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 3066–3073, Apr. 2020.

[9] 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.

[10] 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.

[11] M. A. Daoud, M. W. Mehrez, D. Rayside, and W. W. Melek, "Simultaneous feasible local planning and path-following control for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16358–16370, Sep. 2022.

[12] P. Hang, C. Lv, Y. Xing, C. Huang, and Z. Hu, "Human-like decision making for autonomous driving: A noncooperative game theoretic approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 4, pp. 2076–2087, Apr. 2021.

[13] Q. Liu, X. Li, S. Yuan, and Z. Li, "Decision-making technology for autonomous vehicles: Learning-based methods, applications and future outlook," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 30–37.

[14] F. Leon and M. Gavrilescu, "A review of tracking and trajectory prediction methods for autonomous driving," *Mathematics*, vol. 9, no. 6, p. 660, Mar. 2021.

[15] 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.

[16] 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.

[17] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, 2019, Art. no. eaay7120.

[18] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[19] B. Kim, C. Rudin, and J. A. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[20] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[21] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1803–1831, 2010.

[22] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 589–600, May 2008.

[23] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting CNNs via decision trees," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6261–6270.

[24] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," *Adv. neural Inf. Process. Syst.*, vol. 29, 2016.

[25] S. Bach, A. Binder, K.-R. Müller, and W. Samek, "Controlling explanatory heatmap resolution and semantics via decomposition depth," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2271–2275.

[26] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[27] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2528–2535.

[28] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[29] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 3–19.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[31] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[32] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8827–8836.
[33] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.
[34] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," *ISMIR*, vol. 53, pp. 537–543, Oct. 2017.
[35] M. Hofmarcher, T. Unterthiner, J. Arjona-Medina, G. Klambauer, S. Hochreiter, and B. Nessler, "Visual scene understanding for autonomous driving using semantic segmentation," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. New York, NY, USA: Springer, 2019, pp. 285–296.
[36] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. D. Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 340–341.
[37] Y. Li et al., "A deep learning-based hybrid framework for object detection and recognition in autonomous driving," *IEEE Access*, vol. 8, pp. 194228–194239, 2020.
[38] Y. Shen, S. Jiang, Y. Chen, and K. D. Campbell, "To explain or not to explain: A study on the necessity of explanations for autonomous vehicles," 2020, *arXiv:2006.11684*.
[39] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Towards safe, explainable, and regulated autonomous driving," *CoRR*, vol. abs/2111.10518, pp. 1–8, Nov. 2021.
[40] J. Dong, S. Chen, S. Zong, T. Chen, and S. Labi, "Image transformer for explainable autonomous driving system," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2732–2737.
[41] Y. Sawada and K. Nakamura, "C-SENN: Contrastive self-explaining neural network," 2022, *arXiv:2206.09575*.
[42] Y. Sawada and K. Nakamura, "Concept bottleneck model with additional unsupervised concepts," *IEEE Access*, vol. 10, pp. 41758–41765, 2022.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.
[48] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, pp. 1–4, Jun. 2017.
[49] D. Wang, C. Devin, Q.-Z. Cai, F. Yu, and T. Darrell, "Deep object-centric policies for autonomous driving," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 8853–8859.
[50] P. W. Koh et al., "Concept bottleneck models," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 5338–5348.
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[52] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

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

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

Wei Hua received the Ph.D. degree in applied mathematics from Zhejiang University.

He is currently a Senior Research Expert with the Zhejiang Lab, Hangzhou, China. His current research interests include autonomous driving, intelligent simulation, digital twin, reinforcement learning and AI-based algorithms. His related works have been applied to intelligent transportation, smart city, and smart community.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from the 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 a Research Assistant Professor with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. His current research interests include autonomous driving, robotics and autonomous systems, robotic perception, and autonomous navigation.