IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 December 2022, accepted 4 January 2023, date of publication 10 January 2023, date of current version 13 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3235739

RESEARCH ARTICLE

More Persuasive Explanation Method for End-to-End Driving Models

CHENKAI ZHANG[1], DAISUKE DEGUCHI[1], (Member, IEEE), YUKI OKAFUJI[2], (Member, IEEE), AND HIROSHI MURASE[1], (Life Fellow, IEEE)

[1]Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan
[2]AI Laboratory, CyberAgent Inc., Tokyo 150-6121, Japan

Corresponding author: Chenkai Zhang (zhang.chenkai.d4@s.mail.nagoya-u.ac.jp)

This work was supported in part by the Japan Science and Technology agency (JST) Support for Pioneering Research Initiated by the Next Generation (SPRING) under Grant JPMJSP2125; in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 20K14996; and in part by JST CREST, Japan, under Grant JPMJCR22D1.

ABSTRACT With the rapid development of autonomous driving technology, a variety of high-performance end-to-end driving models (E2EDMs) are being proposed. In order to understand the computational methods of E2EDMs, pixel-level explanations methods are used to obtain the explanations of the E2EDMs. However, little attention has been paid to the excellence of the explanations of E2EDMs. Therefore, in order to build trustworthy E2EDMs, we focus on improving the persuasibility of the explanations of E2EDMs. We propose an object-level explanation method (main approach) for E2EDMs, which masks the objects in the image and then treats the change in the prediction result as the importance of the objects, then we explain the E2EDM by the importance of each object. To further validate the effectiveness of object-level explanations, we propose another approach (validation approach), which trains E2EDMs with object information as input and generates the importance of objects using general explanation methods. Both approaches generate object-level explanations, in order to compare these object-level explanations with traditional pixel-level explanations, we propose experimental methods to measure the persuasibility of explanations of E2EDMs through a subjective and objective method. The subjective method evaluates persuasibility based on the extent to which participants think the importance of features indicated by the explanations is correct. The objective method evaluates the persuasibility based on the human annotation similarity between provided with only the important part of images and provided with the complete images. The experimental results show that the object-level explanations are more persuasive than the traditional pixel-level explanations.

INDEX TERMS Autonomous driving, convolutional neural network, end-to-end model, explainability.

I. INTRODUCTION

The autonomous driving systems can be divided into perception-planning-action pipelines [1] and E2E learning approaches [2]. The perception-planning-action pipeline approaches divide the driving task into smaller sub-modules such as perceiving the environment, planning, making high-level judgments and controlling vehicles. On the other hand, the E2E learning approaches directly learn highly complex transformations that operate on input sensor data and generate end commands. In the field of deep learning, the convolutional neural network (CNN) models are widely used for complex transformations, such as calculating steering/throttle control based on in-vehicle camera images. Due to advances in CNN models, E2E driving systems have better prediction accuracy than the perception-planning-action pipeline driving systems and are gaining more and more attention in the research field.

The perception-planning-action driving systems could provide interpretable explanations by task-specialized sub-module. However, unlike the perception-planning-action driving systems, the E2E driving systems are not inherently interpretable since they simultaneously address intertwined tasks of very different natures: the perception of detecting lanes and objects, the reasoning and planning of the motion of surrounding objects and self-vehicles, and the

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

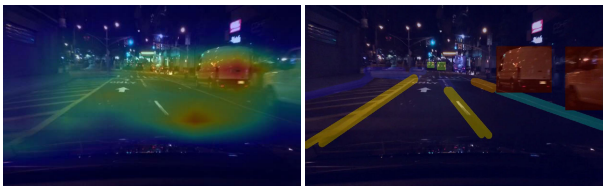
FIGURE 1. Typical scenes in the BDD-3AA dataset. As shown in the upper left image, there are double yellow solid lines on the left, thus the steering left action is not available; there are vacant spaces in the front and right, thus the acceleration and steering right actions are available. The green arrow denotes the corresponding action is available, the red arrow denotes not available.

control of generating driving end-commands. Moreover, the E2E driving systems are unreliable due to their black-box nature. Therefore, as introduced in this survey [3], unlike the perception-planning-action pipeline driving systems, the E2E driving systems still require further exploration of explainability. In this paper, we focus on improving the excellence of the explanations of E2E driving systems. The explanation methods that enable humans to understand the internal processes of E2E driving systems are the prerequisite for E2E driving systems to be accepted by society.

There are many types of explanation methods [3], [4], [5] aimed at explaining E2E models. Among all these methods, attribution-based explanation methods are the most prevalent [6]. In the field of image processing, attribution-based explanations usually refer to saliency maps corresponding to the input images based on the predictions of the model, saliency maps quantify the degree of contribution of each input feature to the predictions. In this paper, our purpose is to evaluate the explanation for the E2EDMs' final decision. Therefore, we believe only the high-level features of E2EDMs contain explanations that indicate the importance of the input feature for the final decision. Moreover, attribution-based approaches are particularly suitable for time-critical tasks, such as driving, where they help users understand E2EDMs only in few seconds.

To make compelling explanations for E2EDMs, a number of problems must be addressed beforehand. First, the majority of the existing driving datasets make it challenging to train E2EDMs in which the relationship between the driving environment and driving actions is unclear to understand; second, whether the current pixel-level explanation method used for explaining the E2EDMs is appropriate for humans is yet to be clarified. Third, despite studies on using object information to improve driving task performance, the impact on E2EDMs explanations has not been investigated.

Previous studies focus on predicting the action made by the driver [7], creating the false impression that only that action was correct. However, there are usually multiple available driving actions to take based on personal preference. Since drivers could randomly choose actions from multiple



Traditional pixel-level explanation | Proposed object-level explanation

FIGURE 2. Ambiguous pixel-level and clear object-level explanation.

available options, the previous driving dataset may confuse driving models with such a problem: "why there are different driving actions for similar driving environments?". To address this problem, we made datasets for a simple driving task where each sample is annotated with the availability of 3 driving actions, acceleration, steering left, and steering right. For such driving tasks, the lane and object information is sufficient for E2EDMs to make the prediction for the availability of 3 driving actions, and we could also depend on the lane and object information to explain the decision made by these E2EDMs. Fig. 1 shows examples of typical scenes and annotation results.

For the current attribution-based explanation methods applied in the E2E models, it is taken for granted to specify the importance of each pixel to explain the calculation methods of the models [6]. On the one hand, this pixel-level expression form explanation is appropriate for the classic object classification tasks, but whether the pixel-level explanation is also compatible with complex tasks such as driving is hardly considered.

This leads us to the evaluation of the explanation, among various properties of explanations introduced in this survey [4], persuasibility is most closely related to humans, that is, to measure the extent to which people understand the explanations. As shown in the heatmap on the left of Fig. 2, people are hard to understand the pixel-level explanation since a huge amount of domains irrelevant to driving are improperly considered to be explanations.

A study on cognitive science [8] suggested that object is the basis of human attention. Recent exploration [9] in E2EDMs also suggests the object plays an important role in high prediction performance, however, the persuasibility influence of objects on the explanation performance has not been investigated. Based on previous research, we consider the object-level expression form explanations are more persuasive, as shown in Fig 2, it is simpler for humans to be persuaded by the object-level explanations, in which only the objects related to driving are given the weight of importance. Therefore, in this paper, we proposed object-level explanations instead of traditional pixel-level explanations to explain E2EDMs by specifying the importance of objects. In order to verify our proposal, we start with a simple driving task where each sample is annotated with the availability of 3 driving actions, acceleration, steering left, and steering right.

In order to explain the CNN models designed for the classic object classification tasks, Zeiler et al. [10] proposed

occlusion-based methods to produce pixel-level explanations by using a gray patch on the image and inspecting how the prediction changes. However, since occlusion-based methods are mainly applied in classical object classification tasks, the potential practicality for making object-level explanations has not been discovered. Therefore, we propose the object-level explanation method (main approach) which covers a gray patch on a particular object bounding box. The difference between our method and Zeiler et al's method is the mask-out area, this difference leads to an entirely different expression form of explanations, where our method could generate object-level explanations designed for the driving tasks, and the previous method could only generate traditional pixel-level explanations designed for the classification tasks. To the best of our knowledge, this is the first study that focuses on the persuasibility difference between object-level explanations and pixel-level explanations. Additionally, our object-level explanation method also satisfies the sensitivity axiom and implementation invariance axiom [11] as an attribution-based explanation method. To further validate the persuasibility influence of object-level explanation, we also train E2EDMs with pure object information as input and directly obtain the importance of objects for the model prediction (validation approach). The object-level explanations made by the validation approach will also be evaluated for persuasibility and compared with traditional pixel-level explanations.

Previous research [6] focused on user satisfaction and decision accuracy when evaluating explanation persuasibility. However, the evaluation method specially designed for driving tasks has not been discovered. To prove that object-level explanations are more persuasive than pixel-level explanations for E2EDMs, we propose a subjective and objective evaluation method to evaluate the persuasibility of different explanations. The subjective method evaluates the persuasibility by the extent to which participants think the importance of features in the explanations is correct. The objective method evaluates the persuasibility by the similarity of human annotation results based only on the important feature from the E2EDM's perspective and based on the complete images.

Through experiments, we have proved that our object-level explanation method could produce more persuasive explanations than traditional pixel-level explanations. We argue that our object-level explanation method is more appropriate to explain E2E models for driving tasks.

The contributions of this paper are as follows:

- We made datasets that driving actions are solely based on the driving environment to help train understandable E2EDMs, and we novelly proposed a method to make an object-based driving dataset.
- We proposed experimental methods to evaluate the explanation persuasibility in driving-related tasks.
- We proposed two approaches to produce object-level explanations. The main approach is an object-level explanation method that could generate persuasive explanations for E2EDMs. The validation approach is proposed to further validate the persuasibility effect of object-level explanations. For the validation approach, we novelly build an object-based driving dataset and train object-based E2EDMs. By calculating the importance of objects using general explanation methods, we could generate object-level explanations from another perspective. Through experiments, both object-level explanations are proved to be more persuasive than the traditional pixel-level explanations in driving tasks.

II. RELATED WORK

In this section, we will briefly review the research on 4 different topics: autonomous driving systems, explainability requirements of E2E driving systems, methods for explaining E2EDMs, and properties of explanation results and evaluation methods for persuasibility. We present these four topics according to the time of their development, the topic that appears first is the reason for the next one to grow.

A. AUTONOMOUS DRIVING SYSTEMS

Autonomous driving systems were strictly designed pipeline systems at the end of the 20th century [12], modular pipeline systems decompose the driving task into several small tasks, involving perceiving the environment, planning, making high-level decisions, and controlling vehicles. Therefore, pipeline systems provide interpretable explanations processing through specialized modules.

However, pipeline systems have several disadvantages. First, they rely on manually selected intermediate representations that are not optimal for driving tasks. Second, they lack flexibility and cannot take into account the uncertainty of the real world. Finally, they easily propagate errors among multiple submodules [13].

To circumvent these problems, people are interested in training E2E driving systems with neural networks. Through a large number of expert data, the E2E driving systems learn a highly complex transition that inputs sensor data and generates end commands (steering angle, throttle) [2]. However, the E2E driving systems are described as black boxes due to lacking transparency compared to pipeline systems, leading us to the explainability needs of E2E driving systems.

B. EXPLAINABILITY NEEDS OF E2E DRIVING SYSTEMS

The need for the explainability of autonomous driving systems depends on the people involved, whether they are end-users, legal authorities, or designers of self-driving vehicles [3]. End-users [14] need to trust autonomous systems before riding. The legal authorities [15] need to obtain systematic explanations for liability, especially in the case of accidents. The designers of self-driving vehicles [16] need to understand the limitations of current models to build better versions.

On the one hand, the E2E driving systems are not inherently interpretable [3], they must simultaneously solve different tasks: perception, planning, decision-making, and control. Therefore, explaining an E2E driving system means decomposing the predictions of each task and making them understandable to humans.

On the other hand, the driving models cannot be fully tested in all cases since it is impossible to list and evaluate every situation that the model may encounter. As a backup solution, this leads us to explain E2EDMs.

C. METHODS FOR EXPLAINING E2E MODELS

There are many surveys about the explanation methods of deep learning, and they classify the explanation methods from different perspectives. Ras et al. [4] classify the explanation methods from the underlying computational methods, and Zablocki et al. [3] classify the explanation methods from the application of autonomous driving technology. In this paper, we classify the explanatory techniques in terms of the type/format of the explanation [5]. We briefly review three major formats as explanations: the rules, the examples, and the attribution.

1) RULE AS EXPLANATION

This category produces a logical rule as the explanation for a trained model. Dhurandhar et al. [17] construct local rule explanations by finding out features that should be minimally present and features that should be minimally absent. E.g., the explanation takes this form *"If an input* $\boldsymbol{x}$ *is classified as the class y, it is because its features* $f_i, \ldots, f_j$ *are present and features* $f_m, \ldots, f_n$ *are absent"*. However, this rule-based explanation method cannot be applied to explain the E2EDM with over-complicated input such as images.

2) EXAMPLE AS EXPLANATION

These methods return other examples as supporting or countering examples to explain an input. The basic intuition is to find the most similar examples considered by the model.

Yeh et al. [18] showed that logit (neurons before softmax) can be decomposed into linear combinations of training point activation in the pre-logit layer. Based on the coefficients of the training points, we can tell whether the similarity with these points is excitatory or inhibitory.

Koh and Liang [19] evaluate the impact of a training instance on the model prediction of an unseen test instance. Firstly, the approximate method is used to calculate the change of model parameters after the training example is changed. Secondly, its effect on the loss of test points can be calculated. By examining the training cases (positive or negative) that have the greatest impact (on the test cases), we can have some understanding of the prediction of the model.

However, the example-based explanation method requires more time to comprehend and hence cannot be applied to time-critical tasks, such as driving tasks.

3) ATTRIBUTION AS EXPLANATION

The Attribution-based explanation method is the attribution of credit or blame to the input features based on their impact on the prediction. The explanation will be a vector with the scores indicating the importance of the input features [20]. Attribution methods can be further divided into three groups: gradient-related methods, occlusion-based methods, and model-agnostic methods.

a: GRADIENT-RELATED METHODS

For image-processing CNNs, the attribution-based explanation is usually represented as a saliency map, a mask of the same size as the input image.

Erhan et al. [21] proposed one of the earliest works on visualization in deep learning models. The activation maximization method is to visualize essential features in any layer of CNNs by optimizing the input feature with the aim that the activation of the chosen unit in any layer is maximized. However, in this paper, we aim to visualize the importance of input features with respect to the final prediction results, thus only the last layer of CNNs contains the necessary information for an explanation since it is used directly for the final prediction.

Simonyan et al. [22] generated the saliency map from the gredients by a single backpropagation pass. There is also Grad-CAM [23], which calculates a saliency map with respect to a certain class on the last convolutional layer, and thus can be universally applied to any CNN models for explanations.

b: OCCLUSION-BASED METHOD

Zeiler et al. [10] covered a gray patch on the image and see how the prediction changes with the different positions covered by the patch: when the patch covers a key area, the prediction performance will be significantly reduced. This method has the advantage of universality to be applied to explain any models with images as input. However, the occlusion-based method has not been used to produce object-level explanations. Therefore, we obtain the object-level explanation by covering a gray patch on a particular object bounding box.

c: THE MODEL AGNOSTIC METHOD

LIME [24] is a well-known approach that can provide attribution-based explanations. Ribeiro et al. approximate black box models with an interpretable model, e.g., the logistics regression model to explain each individual prediction. This method can be universally applied to any model for explanations.

D. PROPERTIES OF EXPLANATION RESULTS AND evaluation METHODS FOR PERSUASIBILITY

Many terms are related to explainability concepts, and several definitions have been proposed for each term, we overview the key concepts related to explainable AI. In human-

computer interaction, Rosenfeld et al. [25] defined explainability as the ability of human users to understand agent logic. According to Doshi-Velez et al. [26], the explanation is a human-understandable description of the process by which a decision-maker takes a specific set of inputs and reaches a specific conclusion. The term explainability often appears with the concept of interpretability. Gilpin et al. [27] used the term interpretability to specify the extent to which an explanation can be understood by humans.

Compared with the vigorous development of various explainable deep learning methods, the progress of evaluation research on these methods is slightly falling behind. For the question "what is a good explanation?", people have put forward different standards as evaluation objectives. Mohseni et al. [28] proposed correctness (or fidelity), correctness requires that the explanations should correctly describe the internal decision-making process. Cui et al. [29] proposed a criterion as coverage, or completeness requires complete information on model explanations to make it repeatable. Yang et al. [30] define generalizability and persuasibility, the former measures the generalization ability of explanations, and the latter is about how well humans comprehend the explanations which is the focus of this paper.

The persuasibility of explanation can be evaluated with the human-annotated ground truth in uncontroversial tasks, such as object detection, which is usually consistent in different user groups. This annotation-based evaluation is generally considered objective because relevant annotations do not change in different user groups. In computer vision tasks, the most common annotations used for persuasibility evaluation are bounding boxes and semantic segmentation. An example can be found in [23], Selvaraju et al. use the boundary box and the metric Intersection over Union (IOU) measure to quantify the performance of persuasibility.

However, in complex tasks, it is not appropriate to use human annotations to evaluate persuasibility, because relevant annotations may not be consistent in different user groups. Therefore, conducting human studies is a common method of assessing the persuasibility of explanations in complex tasks. Lage et al. [31] focused on user satisfaction when evaluating explanation performance, and use human response time and decision accuracy as indicators. Li et al. [6] designed a game that occludes partial explanations and asked users to answer the object name, the persuasibility performance is evaluated based on whether the users correctly answered the name and their subjective satisfaction. However, no evaluation methods designed specifically for driving tasks have been discovered, this is the first research that focuses on persuasibility evaluation in driving tasks.

III. OBJECT-LEVEL EXPLANATION METHOD

To create object-level explanations for a conventional E2EDM using images as input, we present an occlusion-based method. As shown in Fig 3, we will outline a step-by-step process for calculating the importance of each object

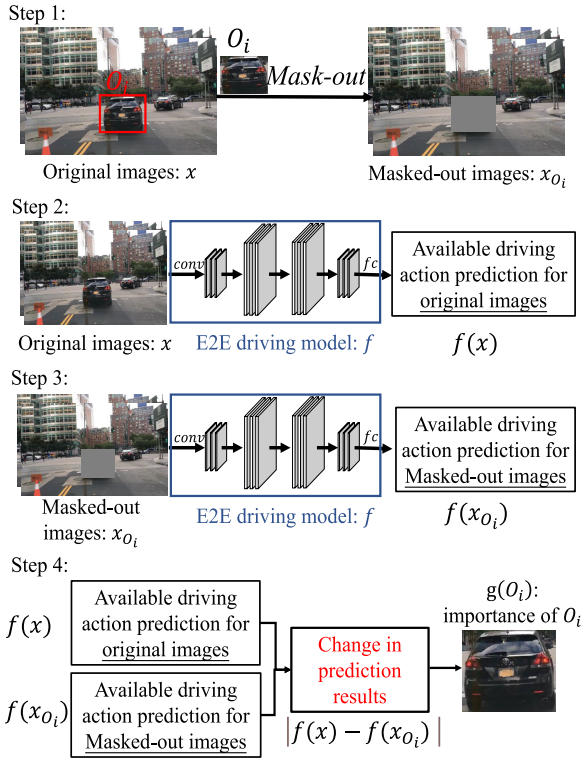


FIGURE 3. The object-level explanation by occlusion-based methods.

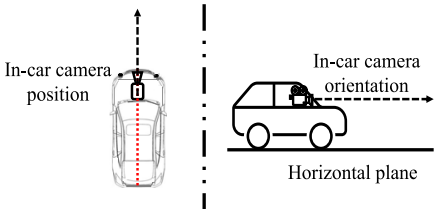


FIGURE 4. In-car camera position and orientation.

in the input images. After gaining the importance score for each object in the input images, we could make heatmaps as object-level explanations for an E2EDM. We display and evaluate these object-level explanations in the VI. EXPERIMENT RESULTS section.

1) As shown in step 1 in Fig. 3, we mask out an object's bounding box in the original images with gray pixels to obtain the mask-out images [10]. We denote the object as O_i, the original images as x, and the mask-out images as x_{O_i}.
2) As shown in step 2 in Fig. 3, we feed the E2EDM the original images and obtain the prediction results. We denote the E2EDM as f and the prediction results for original images as $f(x)$.
3) As shown in step 3 in Fig. 3, we feed the E2EDM the mask-out images and obtain the prediction results. We denote the prediction results for mask-out images as $f(x_{O_i})$.
4) As shown in step 4 in Fig. 3, the significance of the change in predicted results in 2) and 3) is the impor-

TABLE 1. Object category for 3 classes.

Class	Object category
Obstacles	Car, pedestrian, bicycle, bus, truck
Traffic lights	Red traffic light, green traffic light
Lanes	Curb, solid line, dashed line, crosswalk

tance of the object, we denote the importance of the object as $g(O_i)$.

$$g(O_i) = |f(x) - f(x_{O_i})| \tag{1}$$

IV. DATASET

In this section, we first introduce the BDD-3AA (3 available actions) dataset to train 3 pixel-based E2EDMs. We then introduce the object-based BDD-3AA dataset to train 3 object-based E2EDMs.

A. THE BDD-3AA DATASET

1) DATASET LABELS

We consider the classification of actions to be a multi-label classification. Mathematically, given two continuous images in some space X, the goal is to determine the availability of 3 actions, acceleration, steering left, and steering right. This is implemented by mapping $X \mapsto A \in \{0, 1\}^3$. For instance, if the "acceleration" and "steering left" actions are available, then $A = [1, 1, 0]^T$.

2) DATASET IMAGES

There are multi-object tracking videos in the BDD-100K dataset [32]. Each multi-object tracking video has continuous bounding box information for each moveable object. In the BDD-100K dataset, the 51st frame is treated as the key frame of each video, the key frame is also annotated with immovable objects, such as lanes, traffic lights, etc.

In order to build a dataset containing adequate object-level annotation information, we extract the key frame and the previous frame of the key frame from a tracking video. In the key frame, the relative motion of movable obstacles (such as vehicles and pedestrians) could be calculated based on the continuous bounding box information, and the category and position information of immovable objects (such as the lane and traffic lights) is also available. Since the driving task is to predict the availability of 3 actions, acceleration, steering left, and steering right, the lane and object information are sufficient to explain the behavior of such E2EDMs.

We selected images from the scene labeled "city street" and "residential" to build a dataset that focuses on complex driving environments. Considering the huge variation of the orientation and the placement position of the in-car cameras, we manually selected images where the in-car camera is placed in the center of the ego-vehicle, and the orientation of the in-car camera is parallel to the horizontal plane and the central axis of the vehicle, as shown in Fig. 4. We resulted in a 500 2-frame video clips BDD-3AA dataset.

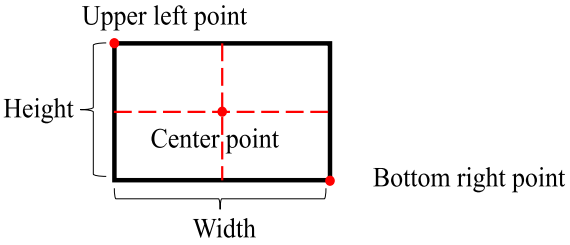


FIGURE 5. The calculation of width, height, and center position based on the bounding box coordinate information.

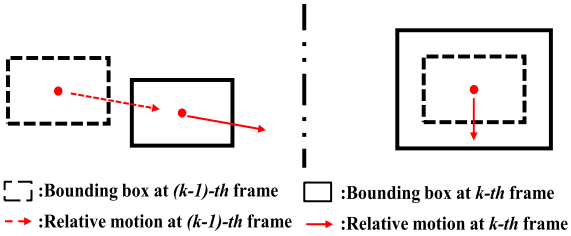


FIGURE 6. The relative motion calculation method based on the center point position change, and the area change between bounding boxes at *(k-1)-th* and *k-th* frame.

B. OBJECT-BASED BDD-3AA DATASET

In order to train object-based driving models, we transfer the BDD-3AA dataset to the object-based BDD-3AA dataset.

We classify the objects that are closely related to the driving tasks, we divide them into three classes: obstacles, traffic lights, and lanes. For each class, the category of objects is shown in Table 1.

In addition to category information, for lanes, we need to know their position information in the 2D image. For traffic lights, we need to know their size and position information in the 2D image. For obstacles, we need to know their relative motion to the ego-vehicle, their size, and position information in the 2D image. Since the driving task is to predict the availability of 3 actions, acceleration, steering left, and steering right, the lane and object information are sufficient to train such E2EDMs.

According to the coordinate information of the bounding box, we calculate the width, height, and center position of the bounding box (as shown in Fig. 5). We denote the k as the serial number of the key frame, the CP_k, W_k, and H_k as the center point, width, and height information of the object bounding box at the k-th frame. The $CP_{(k,x)}$ and $CP_{(k,y)}$ are the center point horizontal and longitudinal coordinates of the object bounding box in the k-th frame.

The bounding box of obstacles becomes bigger in the image as they get closer to the ego-vehicle, therefore, we approximate their relative motion (RM) to the ego-vehicle based on the *center point* position change, and the area change (AC) between bounding boxes in two adjacent frames (as shown in Fig. 6). As shown in Eq. (2), in order to calculate relative motions with similar magnitude for close-range vehicles and long-range vehicles based on area change, we use the ratio between the area change and the area of the *(k-1)-th* frame obstacle for normalization. The constant C to adjust the

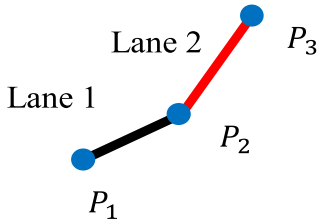


FIGURE 7. We decompose a curvy lane into several straight lanes. We use the coordinate information of P_1 (start point) and P_2 (end point) to represent the black lane (lane 1), and use P_2 (start point) and P_3 (end point) to represent the red lane (lane 2).

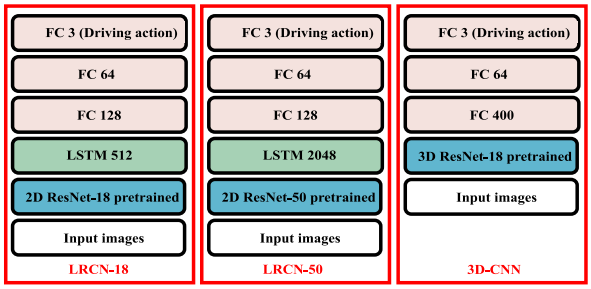


FIGURE 8. The architectures of pixel-based E2EDMs.

importance of the area change is set to 5 in Eq. (4).

$$AC = \frac{W_k * H_k - W_{k-1} * H_{k-1}}{W_{k-1} * H_{k-1}} \tag{2}$$

$$RM_x = CP_{(k,x)} - CP_{(k-1,x)} \tag{3}$$

$$RM_y = CP_{(k,y)} - CP_{(k-1,y)} + C * AC \tag{4}$$

As shown in Fig. 7, to be able to simply represent the curvy lanes, we decompose each curvy lane into several straight lanes and record the coordinate of the start and end points of each straight lane. We denote the *SP* as the start point and the *EP* as the end point of each straight lane.

For each object class, we use different vectors to represent their information. We use the ($category$, $CP_{(k,x)}$, $CP_{(k,y)}$, W_k, H_k, RM_x, RM_y) to represent each obstacle, the ($category$, $CP_{(k,x)}$, $CP_{(k,y)}$, W_k, H_k) to represent each traffic light, the ($category$, $SP_{(k,x)}$, $SP_{(k,y)}$,$EP_{(k,x)}$, $EP_{(k,y)}$) to represent each straight lane.

In order to obtain the features of all objects while maintaining the features of various driving environments with the same length, we utilize the *bag of words model* and *K-means clustering* for the data processing. The details are introduced in APPENDIX A.

V. EXPERIMENT

A. PIXEL-BASED E2EDMs

To build a pixel-based E2EDM with images as input, we apply the Long-term Recurrent Convolutional Network (LRCN) [33] and 3D Convolutional Neural Networks (3D CNN) [34] for spatiotemporal neural networks.

The LRCN model investigates the spatiotemporal tasks by applying a long short-term memory (LSTM) recurrent neural network to the output of a CNN.

The 3D CNN is similar to 2D convolutional networks, in addition to height and width, 3D CNN also has the third dimension depth (temporal). Instead of having a 2D filter moving within the image along height and width, now we have a 3D filter moving along with height, width, and depth.

We fine-tune two LRCN networks with Resnet-18 and Resnet-50 backbones on the BDD-3AA dataset. The Resnet-18 and Resnet-50 backbones are pretrained on ImageNet [35]. As shown in Fig. 8, each backbone is connected to a LSTM, and then a stack of fully connected layers. We call these two LRCN networks the LRCN-18 and LRCN-50.

We fine-tune a 3D CNN network with an 18-layer Resnet3D backbone on the BDD-3AA dataset. The Resnet3D backbone is pretrained on Kinetics [34], and connected to a stack of fully connected layers. The architecture can also be seen in Fig. 8, we call this network 3D-CNN.

All fully connected layers used ReLU as their activation.

We generate explanations for the above three pixel-based E2EDMs by a pixel-level explanation method and an object-level explanation method, respectively, these 6 explanations will be introduced in subsection C. 1) and 2) for the upcoming persuasibility evaluation.

B. OBJECT-BASED E2EDMs

We trained the object-based E2EDMs in an E2E training fashion. The input of these models is a 168-length feature vector representing driving-related object information, and the output is the available driving actions.

The three machine learning models are the logistics regression model (LR), the random forest model (RF), and the multilayer perceptron model (MLP). All model's hyperparameters are manually decided based on prior experience.

All models are trained with the scikit-learn library. The hyperparameters of LR are $C = 1.0$ as the regularization strength and the $solver = liblinear$ as the optimization algorithm. The hyperparameters of RF are $n-estimators = 29$ as the number of trees in the forest, and the $max-depth = 21$ as the maximum depth of the tree. The architecture of MLP is 3 hidden layers with (90, 190, 90) neurons respectively. The other hyperparameters are set by default.

We generate explanations for the above three object-based E2EDMs by an object-level explanation method, these 3 explanations will be introduced in subsection C. 3) for the upcoming persuasibility evaluation.

C. 9 EXPLANATIONS FOR E2EDMs

In order to analyze the effect of object-level explanation, we made 9 different explanations for persuasibility evaluation. We briefly introduce each explanation method below.

1) PIXEL-LEVEL EXPLANATIONS FOR THE PIXEL-BASED MODEL (TRADITIONAL APPROACH)

For the pixel-level explanation method, we adopt the Grad-CAM [23] method for each model, we call the pixel-level

explanation generated from LRCN-18, LRCN-50, and 3D-CNN models as LRCN-18-P, LRCN-50-P, 3D-CNN-P.

2) OBJECT-LEVEL EXPLANATIONS FOR THE PIXEL-BASED MODEL (MAIN APPROACH)

For the pixel-based E2EDMs, we apply our method introduced in section III to obtain the importance of each driving-related object, we call the object-level explanations generated from LRCN-18, LRCN-50, 3D-CNN models as LRCN-18-O, LRCN-50-O, 3D-CNN-O.

3) OBJECT-LEVEL EXPLANATIONS FOR THE OBJECT-BASED MODEL (VALIDATION APPROACH)

In order to further verify the persuasibility effect of object-level explanation, we trained E2EDMs using pure object information as input and used LIME [24] to provide an explanation that is automatically at the object level. We call the object-level explanations generated from LR, RF, and MLP models as LR-O, RF-O, and MLP-O.

D. IMPLEMENTATION DETAIL

All models are trained based on the BDD-3AA or the object-level BDD-3AA dataset. For the BDD-3AA dataset, each video clip that contains two continuous images is used, the input size of images is 1280×720. For the object-level BDD-3AA dataset, a 168-length feature to represent each video clip is used. Each dataset is divided into a training set of 300 examples, a validation set of 100 examples, and a test set of 100 examples. To evaluate the prediction accuracy of each E2EDMs, we apply 5-fold cross-validation to train each model five times on the different training datasets and evaluate the average accuracy on corresponding test datasets.

The dataset is imbalanced, i.e., most of the ''acceleration'' actions are available but most of the ''steer left'' and ''steer right'' actions are not, thus we use the macro F1 score to evaluate the prediction accuracy of driving models by calculating the average value of F1 score of three actions.

$$F1_{Macro} = \frac{F1(\hat{A}_a, A_a) + F1(\hat{A}_l, A_l) + F1(\hat{A}_r, A_r)}{3}, \quad (5)$$

where $\hat{A}$ is the prediction action and A is the ground truth action. A_a is the acceleration action, the A_l is the steer left action, and the A_r is the steer right action.

E. THE PERSUASIBILITY EVALUATION METHOD

We have the pixel-level explanations and object-level explanations results. For the pixel-level explanation, we generate 3 explanations for 3 pixel-based driving models. For the object-level explanation, we generate 6 explanations for 3 pixel-based and 3 object-based driving models. Based on the experimental method proposed in [6] and [31], we introduce the evaluation method to analyze the influence of object-level explanations on persuasibility.

Our central idea to evaluate persuasibility is to find out whether the human judgment on the importance of input features is the same as the E2EDM. We use two experimental methods to evaluate the similarity of two judgments, the objective evaluation method, and the subjective evaluation method. The subjective evaluation approach could directly measure how well participants agreed with E2EDM-generated explanations, whereas the objective evaluation method could rigorously measure the persuasibility from the perspective of driving actions. Moreover, the subjective persuasibility score is easily influenced by participants' prior experience with the explanations previously displayed, whereas the objective persuasibility score could be utilized in the future to compare with other explanations.

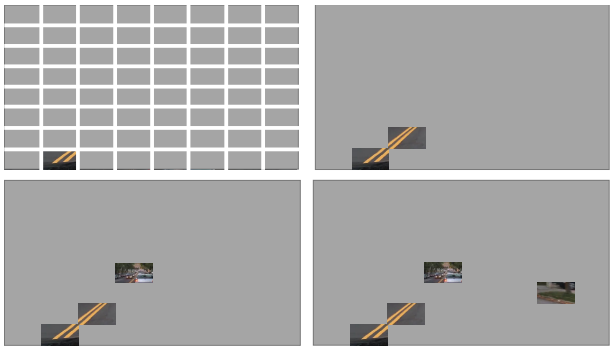


FIGURE 9. Gradually display the important parts of an image, the original image is the upper left image in Fig 1.

1) THE OBJECTIVE EVALUATION METHOD

In the objective experiment, we only show the important part of an image to participants according to the explanations, if the participants can make the same prediction results based on a partially shown image as when they see the complete image, then it means the explanations can correctly extract the driving-related features that are considered informative by participants, i.e., the explanations are persuasive.

We cut an image into grids of the same size. As shown in Fig. 9, for the image of 1280×720, we can divide it into 8×8 grids, the gray area means that there is no information related to driving. The participants are asked to predict the available driving action only based on the shown parts, and the gray area should be considered as an area with nothing.

1) We show the most important grid in the E2EDM's view, and the participants judge the availability of the driving actions only based on the shown grid (upper left in Fig. 9).
2) We show the first and second important grids at the same time, the participants judge the driving actions based on the two shown grids (upper right in Fig. 9).
3) We show the first, second, and third important grids at the same time, the participants judge the driving actions based on the three shown grids (lower left in Fig. 9).
4) We show the first, second, third, and fourth important grids at the same time, the participants judge the driving actions based on the four shown grids (lower right in Fig. 9).

TABLE 2. The prediction accuracy for each driving model.

	Pixel-based E2EDM			Object-based E2EDM		
	LCRN-18	LCRN-50	3D-CNN	LR	RF	MLP
F1-score	**76.26%**	73.35%	75.08%	71.39%	70.72%	71.68%

TABLE 3. The objective and subjective persuasibility experimental results for pixel-level explanations and object-level explanations from our main approach. The introduction for each explanation is shown in V. Experiment, subsection C. The higher the objective F1-score and the subjective score, the more persuasive the explanation is.

		Pixel-level explanation (traditional approach)	Object-level explanation (our main approach)
Objective F1-score	LCRN-18	73.68%	**74.55%**
	LCRN-50	71.89%	**75.14%**
	3D-CNN	73.93%	**75.74%**
Subjective score	LCRN-18	3.25	**4.20**
	LCRN-50	2.91	**3.52**
	3D-CNN	3.20	**3.82**

TABLE 4. The objective and subjective persuasibility experimental results for object-level explanations from our main approach and object-level explanations from our validation approach. The introduction for each explanation is shown in V. Experiment, subsection C. The higher the objective F1-score and the subjective score, the more persuasive the explanation is.

	Object-level explanation (ours)					
	Main approach			Validation approach		
	LCRN-18-O	LCRN-50-O	3D-CNN-O	LR-O	RF-O	MLP-O
Objective F1-score	74.55%	75.14%	75.74%	76.26%	76%	75.42%
	Average: 75.14%			Average: **75.89%**		
Subjective score	4.20	3.52	3.82	3.99	3.85	3.79
	Average: 3.84			Average: **3.87**		

We have two explanations with different expression forms, pixel-level, and object-level explanations. The two different expression form explanations require different methods of calculating the importance of grids. For the pixel-level explanation, we calculate the sum of the importance of all pixels in each grid as the importance score of the grid. For the object-level explanation, we denote the importance of the most important object in each grid as the importance of the grid. We consider the object belongs to a grid provided the center point is located in that grid.

Finally, we recorded the driving action annotation for the complete image by the participants, which was used as the ground truth for the previously collected annotation results.

2) THE SUBJECTIVE EVALUATION METHOD

We show a video clip and a heatmap to the participants, the heatmap is made based on the explanations to indicate the importance of each input feature. The participants score the heatmap from "1" to "5", with "1" being the heatmap is hard to trust and "5" being the heatmap is easy to trust.

3) PARTICIPANTS

We recruited 8 subjects for our experiments. All the participants have driver's licenses. Participants were given a tutorial on each task and the interface.

VI. EXPERIMENT RESULTS

A. THE ACCURACY OF DRIVING MODELS

The prediction accuracy for each E2EDM is shown in Table 2, we can see traditional pixel-based E2EDMs have higher prediction accuracy than object-based E2EDMs. We believe the main reason is that pixel-based E2EDMs could extract more features for predictions due to the input images which contain rich information and complex network architecture.

B. PERSUASIBILITY EVALUATION RESULTS

1) OBJECTIVE PERSUASIBILITY EVALUATION RESULTS

In the objective evaluation method, we collected the action labels in the case of showing only the important parts of the images and the complete images. We used the macro F1 score again to measure the similarity between the action labels of the partially displayed images and the complete images. The higher the score, the more abundant the driving-related information is in the partially displayed images, i.e., the more persuasive the explanations are.

As shown in Table 3, we present the F1 score as an objective indicator for the persuasibility of the 3 pixel-level (LCRN-18-P, LCRN-50-P, and 3D-CNN-P) and 3 object-level explanations from our main approach (LCRN-18-O, LCRN-50-O, and 3D-CNN-O). We can see that the object-level explanations from our main approach are all better than the pixel-level explanations for each model and each score.

As shown in Table 4, we present the F1 score as an objective indicator for the persuasibility of the 3 object-level explanations from our main approach (LCRN-18-O, LCRN-50-O, and 3D-CNN-O) and 3 object-level explanations from our validation approach (LR-O, RF-O, and MLP-O). Averagely speaking, We can see that the validation approach is slightly better than our main approach.

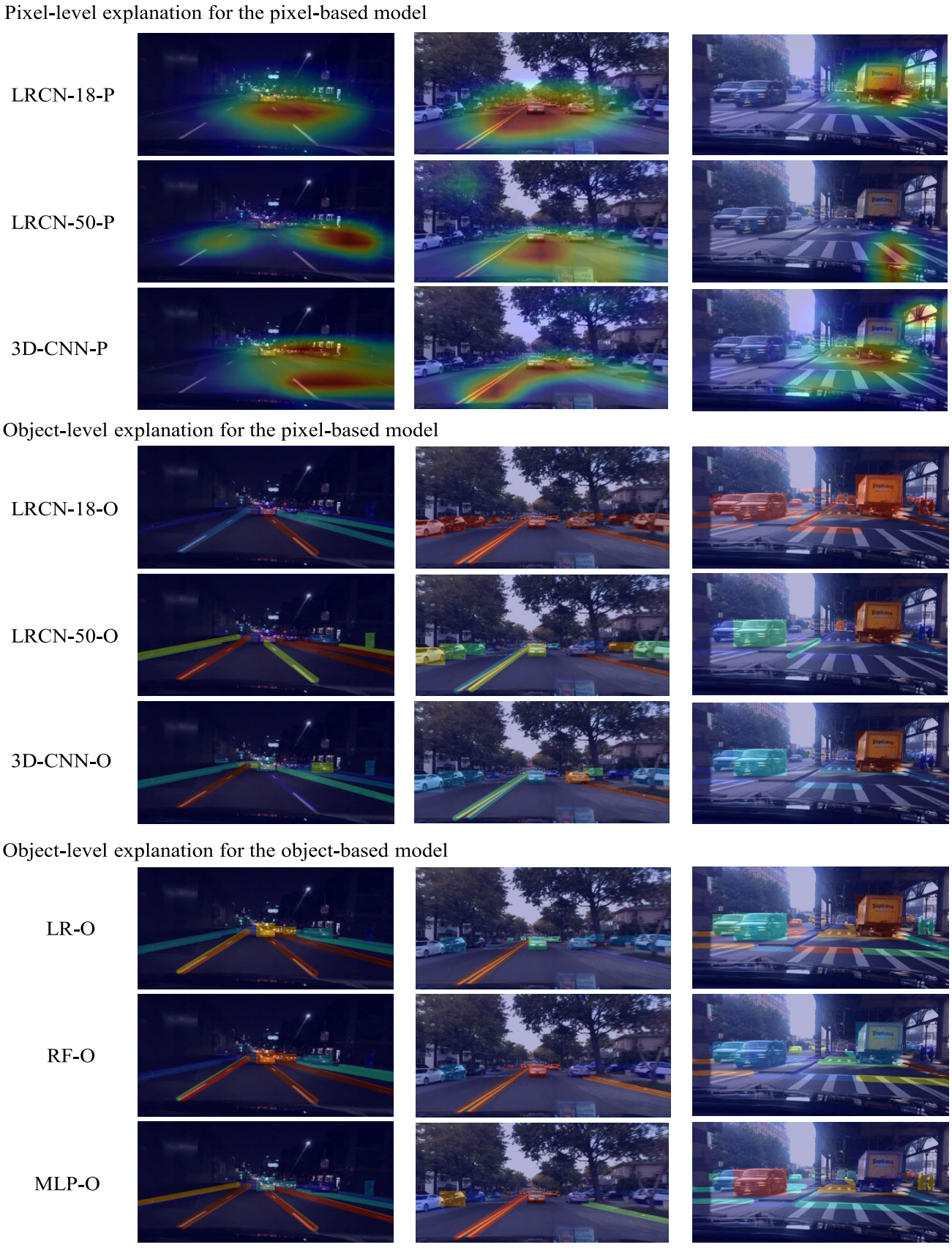


FIGURE 10. Heatmaps for each explanation are shown in the subjective evaluation. Warmer colors indicate higher importance and colder colors indicate lower importance. Red is the warmest color and purple is the coldest in these maps.

2) SUBJECTIVE PERSUASIBILITY EVALUATION RESULTS

As shown in Fig. 10, there are examples in the subjective evaluation experiment for 3 pixel-level explanations and 6 object-level explanations. As shown in Table 3, we present the subjective score as a subjective indicator for the persuasibility of the 3 pixel-level (LCRN-18-P, LCRN-50-P, and 3D-CNN-P) and 3 object-level explanations from our main approach (LCRN-18-O, LCRN-50-O, and 3D-CNN-O). The participants score the heatmap with "1" meaning the heatmap is hard to trust and "5" meaning the heatmap is easy to trust, hence bigger subjective scores indicate more persuasive explanations. We can see that the object-level explanations from our main approach are better than the pixel-level explanations.

As shown in Table 4, we present the subjective score as a subjective indicator for the persuasibility of the 3 object-level explanations from our main approach (LCRN-18-O, LCRN-50-O, and 3D-CNN-O) and 3 object-level explanations from our validation approach (LR-O, RF-O, and MLP-O). Averagely speaking, We can see that the validation approach is slightly better than our main approach.

3) DISCUSSION

We used an occlusion-based method (main approach) for the pixel-based E2EDMs to generate object-level explanations (LCRN-18-O, LCRN-50-O, and 3D-CNN-O). As shown in Table 3, despite being generated by the same pixel-based E2EDMs, these object-level explanations are more persuasive than pixel-level explanations. One possible reason is that the closer the object is to the ego-vehicle, the larger the object bounding box shown in the image, and generally speaking, the more important the object is for the driving task. The occlusion-based method is masking the bounding box of an object, and then inputting the occluded image into the E2EDM, the importance of the object is the degree of change in the prediction result. Such an explanation method has the property that the larger the masked area is, the more the prediction result of the model changes, regardless of the object category.

To further validate the persuasibility effect of object-level explanations, we eliminate the effect of the bounding box size factor by proposing a validation approach to generate object-level explanations. We train simple machine learning models with object information as input and obtain an object-level explanation for persuasibility evaluation (LR-O, RF-O, and MLP-O). As shown in Table 4, averagely speaking, the object-level explanation generated by the object-based driving model is slightly better than the object-level explanation generated by the pixel-level E2EDMs (LCRN-18-O, LCRN-50-O, and 3D-CNN-O). We believe the experimental results indicate that pure object information could train even more persuasive E2EDMs.

Despite the large gap in prediction accuracy between the object-based E2EDMs and the pixel-based E2EDMs (shown in Table 2), we can observe that the explanation results given by the object-based E2EDMs are comparable or even more persuasive than the pixel-based E2EDMs (shown in Table 3 and 4). This demonstrates that, while numerous features can be derived from the image to assist the pixel-based driving models in attaining higher prediction accuracy, some of the features are incomprehensible and unacceptable to people, hence the explainability of pixel-based E2EDMs is relatively lower. We believe that this is also a reflection of the accuracy-explainability trade-off in the deep learning field.

VII. CONCLUSION

In this paper, in order to improve the excellence of the explanations of pixel-based E2EDMs, we proposed the main approach to explain the pixel-based E2EDMs by generating object-level explanations. In addition, in order to validate the persuasibility effect of object-level explanations, we proposed a validation approach to generate object-level explanations by training and explaining object-based E2EDMs. Finally, we proposed experimental methods to measure the explanation's persuasibility. The code of our work will soon be open source.

Through experiments, object-level explanations generated by both approaches are more persuasive than pixel-level explanations. Compare to the previous research, we challenged the most widely used explanation method and we proved that object-level explanations are easier than pixel-level explanations for people to understand E2EDMs. We intend to make the explanations more persuasive in the future by making them succinct in order to avoid confusion caused by overly detailed explanations.

Ensuring the explanations of E2EDMs are persuasive is the first step in ensuring that they are trustworthy. In this paper, we demonstrate that object-level explanations are easier for humans to understand, but the question of whether the E2EDMs themselves are trustworthy has not been addressed, we leave this to future work. In addition, in order to make explanations that concern more objects, we plan to apply our object-level explanation method to full E2EDMs by adding more categories of objects, such as fences, trees, trash cans, etc. to the dataset annotations and adding extra information for each object such as the height, and orientation information to make the object have 3D properties.

APPENDIX A TRANSFORM OBJECTS INFORMATION TO A FIXED-LENGTH FEATURE

We introduce the details to obtain the features of all objects while maintaining the features of various driving environments with the same length. Since the number and category of objects appearing in video clips are various, we adopt the well-known *bag of words model* and *K-means clustering* to transform all object information in a video clip into a fixed-length feature.

There are different object information, such as size, position, and relative motion. We perform different *K-means clustering* methods for each piece of information. We divide the size information into 2 clusters, representing the object is

far or *near* from the ego-vehicle, the position information into 3 clusters, representing the object on the *left*, *front*, or *right side* with regards to the ego-vehicle, the relative motion information into 4 clusters, representing the object's relative motion direction as *forward*, *backward*, *go left*, or *go right* regards to the ego-vehicle. Different categories of objects must be grouped by different *K-means clustering*. E.g., the *K-means clustering* for the size information of a pedestrian is different from the *K-means clustering* for the size information of a car.

We perform different *K-means clustering* for each object information of an object. E.g., for all possible types of a pedestrian, after arranging and combining the size, position, and relative motion information together, there are $2*3*4 = 24$ different types. Therefore, we use a feature with a length of 24 to describe whether each type of pedestrian exists in a video clip image. Each obstacle is the same as the pedestrian, e.g., a car needs 24-length features to be represented as well. By that analogy, each lane needs $3 * 3 = 9$ length features for the start point and end point position information, and each traffic light needs $2 * 3 = 6$ length features for size and position information.

In each feature describing an object, each digit could be 1 or 0, it represents whether a type of object exists in this video clip and ignores the quantity. E.g., for a feature "100000" for the green traffic light, "100000" could mean there are *far* and on the *left side* green traffic lights exist in this clip.

According to the method described above, for each video clip, we transform the information of all driving-related objects to a feature with the length of $24*5+6*2+9*4 = 168$.

ACKNOWLEDGMENT

The author Chenkai Zhang would like to thank the Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System.

REFERENCES

[1] J. Levinson, J. Askeland, J. Becker, J. Dolson, D. Held, S. Kammel, J. Z. Kolter, D. Langer, O. Pink, V. Pratt, M. Sokolsky, G. Stanek, D. Stavens, A. Teichman, M. Werling, and S. Thrun, "Towards fully autonomous driving: Systems and algorithms," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 163–168.

[2] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Müller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.

[3] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," 2021, *arXiv:2101.05307*.

[4] G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *J. Artif. Intell. Res.*, vol. 73, pp. 329–397, Jan. 2022.

[5] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021.

[6] J. Li, D. Lin, Y. Wang, G. Xu, and C. Ding, "Towards a reliable evaluation of local interpretation methods," *Appl. Sci.*, vol. 11, no. 6, p. 2732, Mar. 2021.

[7] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2174–2182.

[8] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, nos. 1–2, pp. 1–46, 2001.

[9] Y. Xu, X. Yang, L. Gong, H.-C. Lin, T.-Y. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9523–9532.

[10] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.

[11] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[12] C. Urmson et al., "Autonomous driving in urban environments: Boss and the urban challenge," *J. Field Robot.*, vol. 25, pp. 425–466, Jan. 2008.

[13] R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, and A. Weller, "Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 1–9.

[14] J. K. Choi and Y. G. Ji, "Investigating the importance of trust on adopting an autonomous vehicle," *Int. J. Hum.-Comput. Interact.*, vol. 31, no. 10, pp. 692–702, 2015.

[15] S. Rathi, "Generating counterfactual and contrastive explanations using SHAP," 2019, *arXiv:1906.09293*.

[16] Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: Automated testing of deep-neural-network-driven autonomous cars," in *Proc. 40th Int. Conf. Softw. Eng.*, May 2018, pp. 303–314.

[17] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–15.

[18] C. K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar, "Representer point selection for explaining deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–13.

[19] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1885–1894.

[20] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018.

[21] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, p. 1, 2009.

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

[23] 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," *Int. J. Comput. Vis.*, vol. 2017, pp. 618–626, Dec. 2017.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144.

[25] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auto. Agents Multi-Agent Syst.*, vol. 33, pp. 673–705, Jan. 2019.

[26] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, A. Weller, and A. Wood, "Accountability of AI under the law: The role of explanation," 2017, *arXiv:1711.01134*.

[27] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2018, pp. 80–89.

[28] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," 2018, *arXiv:1811.11839*.

[29] X. Cui, J. M. Lee, and J. Hsieh, "An integrative 3C evaluation framework for explainable artificial intelligence," in *Proc. AI Semantic Technol. Intell. Inf. Syst. (SIGODIS)*. Cancún, Mexico: AIS eLibrary, 2019, pp. 1–10.

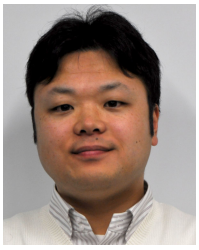
[30] F. Yang, M. Du, and X. Hu, "Evaluating explanation without ground truth in interpretable machine learning," 2019, *arXiv:1907.06831*.

[31] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. Gershman, and F. Doshi-Velez, "An evaluation of the human-interpretability of explanation," 2019, *arXiv:1902.00006*.

[32] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.

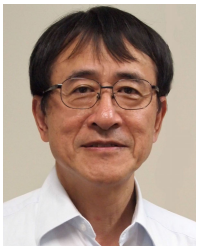
[33] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
[34] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
[35] 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.

CHENKAI ZHANG received the B.Eng. and B.A. degrees from the Dalian University of Technology, Dalian, China, in 2019, and the B.Eng. and M.Eng. degrees from Ritsumeikan University, Shiga, Japan, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree in information science with Nagoya University, Japan. His main research interests include explainable artificial intelligence and the reliability of automatic driving.



DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow at Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor at the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor at Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. His research interests include the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.

YUKI OKAFUJI (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in engineering from Kobe University, in 2014, 2015, and 2018, respectively. From 2018 to 2019, he was a Research Fellow of the Japan Society for the Promotion of Science (DC2 and PD), Kobe University, and Ritsumeikan University. From 2019 to 2022, he was an Assistant Professor at Ritsumeikan University. In 2017, he was a Visiting Researcher with the University of Leeds. Since 2022, he has been a Research Scientist with the AI Laboratory, CyberAgent Inc., and a Visiting Associate Professor with Ritsumeikan University. His research interests include human behavior analysis and human–robot interaction.



HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined at Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist at Columbia University, NY, USA. He has been a Professor with Nagoya University, since 2003. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of IPSJ and IEICE. He received the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.