Toward Explainable End-to-End Driving Models via Simplified Objectification Constraints

Chenkai Zhang, Daisuke Deguchi, *Member, IEEE*, Jialei Chen, and Hiroshi Murase, *Life Fellow, IEEE*

***Abstract*— The end-to-end driving models (E2EDMs) convert environmental information into driving actions using a complex transformation which makes E2EDMs have high prediction accuracy. Due to the black-box nature of transformation, the E2EDMs have low explainability. To solve this problem, explanation methods are used to generate explanations for observation. Based on current explanation methods, previous studies tried to further improve the explainability of E2EDMs by integrating an object detection module, however, these methods have many problems: Firstly, due to the requirement of the object detection module, they lack flexibility. Secondly, they neglect an essential property, *i.e.*, simplicity, to improve explainability. In this paper, since humans prefer object-level and simple explanations in driving tasks, we argue that explainability is decided by two properties which are the objectification degree (the extent to which driving related-object features are utilized) and simplification degree (the simplicity of the explanation), thus we propose Simplified Objectification Branches (SOB) to improve the explainability of E2EDMs. Firstly, this structure could be integrated into any existing E2EDMs and thus have high flexibility. Secondly, the SOB explicitly improves the simplification degree without sacrificing the objectification degree of the explanations. By designing several indicators, *i.e.*, heatmap satisfaction, driving action reproduction score, deception level, *etc.*, we proved that SOB could help E2EDMs generate better explanations. Notably, the SOB could also further enhance E2EDMs' prediction accuracy.**

***Index Terms*— Explainability, autonomous vehicles, deep learning, convolutional neural networks.**

I. Introduction

AUTONOMOUS driving systems are closely related to human safety and ensuring that these systems are reliable is important. Specifically, if the system's driving decisions differ from what humans consider reasonable, humans are entitled to request an explanation of the system's driving decisions. In addition, previous research has shown that there are two indispensable components to human trust in a model: performance-based trust and process-based trust [1]. The former corresponds to the model's prediction accuracy, and the latter corresponds to the model's explainability, where previous research has defined explainability as the extent to which a model's predictions can be understood by humans [2].

The autonomous driving models can be divided into two types [3]: modular driving models [4], [5] and end-to-end driving models (E2EDMs) [6], [7], [8], [9]. Modular driving models are designed based on the human driving strategy, *i.e.*, the perception-planing-action pipeline, thus modular driving models are interpretable. The definition of interpretable is that the model's predictions can be understood by observing the modelitself [10], [11], [12]. However, the modular driving models select hand-craft features that are not optimal for the tasks [5], [9], *i.e.*, they have the drawback of low prediction accuracy. To make driving models that have higher prediction accuracy, E2EDMs [6], [7], [8], [9] are developed by using a complex transformation to convert environmental information into driving actions. This complex transformation can learn optimal features that fit the current task, making E2EDMs have high prediction accuracy. However, E2EDMs have poor explainability due to the black-box nature of the complex transformation. Since the perception-planing-action pipeline architecture is prone to error propagation and accumulation, in order to make driving models that have both high prediction accuracy and explainability, researchers tend to solve the explainability issue of E2EDMs instead of the prediction accuracy issue of modular driving models.

To explain E2EDMs, explanation methods are employed to generate explanations for future observations [10], [11], [13]. There are two fundamental properties of explanations, persuasibility and fidelity. Persuasibility represents how well people understand and agree with the explanations. Fidelity represents whether the explanation can faithfully reflect the model's computational method. For fidelity, current explanation methods can be divided into two categories [14]: 1. Passive explanation methods, 2. Active explanation methods. Passive explanation methods are applied after training models, they do not intervene in the model's architecture or training process; instead, they analyze the model's outputs, internal features, and weights. On the other hand, active explanation methods are considered during the model's design and training process. These methods introduce components that generate explanations and use these explanations for computing the final prediction, thus explanations generated by active explanation methods are faithful. In this paper, to ensure the generated explanations are faithful, we use active explanation methods to explain our proposal and various baselines.

There are textual-based explanation methods [15], [16], [17], [18] and visual-based explanations [19], [20], [21], [22],

Manuscript received 12 September 2023; revised 21 February 2024; accepted 3 April 2024. 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 23H03474, and in part by JST Core Research for Evolutional Science and Technology (CREST) under Grant JPMJCR22D1. The Associate Editor for this article was S. Kumari. *(Corresponding author: Chenkai Zhang.)*

The authors are with the Graduate School of Informatics, Nagoya University, Nagoya 464-8601, Japan (e-mail: zhang.chenkai.d4@s.mail.nagoya-u.ac.jp).

Digital Object Identifier 10.1109/TITS.2024.3385754

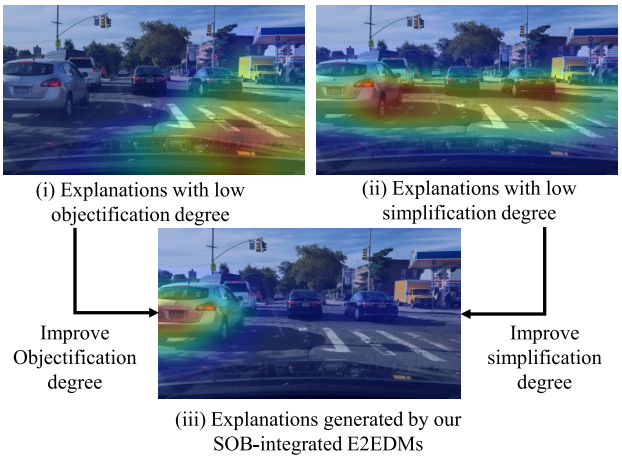

Fig. 1. (i) is the explanation generated by a traditional E2EDM, which suffers from a low objectification degree (the extent to which driving-related object features are utilized); (ii) is the explanation generated by a ROB-integrated E2EDM [29], despite it could have a high objectification degree, it suffers from a low simplification degree (the simplicity of the explanation), leading to overcomplex explanations tend to confuse and deceive humans, *i.e.*, humans are unable to recognize the precise cause responsible for the prediction; (iii) is the explanation generated by our proposed SOB-integrated E2EDM, which has higher simplification and objectification degrees that could improve the explainability of E2EDMs.

[23], [24] methods, the former generates natural language to explain why the driving models perform a specific driving action, the latter uses visual information, i.e., images to offer intuitive explanations. Compared to textual-based explanations, visual-based explanations have the advantage at time-critical tasks, such as driving, thus in this paper, we focus on visual-based explanations. Among various visual-based explanation methods, attribution-based methods [10], [12], [13] are widely utilized to calculate the importance score of each input element in the model's prediction. As shown in Fig. 1, the heatmaps illustrate the importance of pixels in the prediction results, serving as explanations. In driving tasks, these explanations are particularly suitable as they facilitate a quick understanding of the predictions [13]. Additionally, since the basis of human attention lies in objects [25], as shown in Fig. 1 (i), the explanation with a low objectification degree (the extent to which driving-related object features are utilized) is less persuasive [19].

Therefore, the previous studies to enhance the explainability of E2EDMs focus on improving the explanations' objectification degree by integrating an object detection module into the driving model [20], [21], [22], [23], [24]. This enables the model to capture specific object elements and generate precise object-level explanations. However, due to the specific structural requirements of an object detection module, this approach has limited flexibility. In addition, they neglect the importance of simplicity in explanations, leading to two critical issues. Firstly, prior research [26], [27], [28] has demonstrated that complex explanations tend to confuse humans and undermine the overall explainability of E2EDMs. Secondly, complex explanations can also deceive humans. As illustrated in Fig. 1 (ii), the explanation suggests that this particular E2EDM relies on numerous elements to make predictions, which might appear convincing to humans, especially because the most important vehicle (the one close to us on the left) is highlighted. However, in reality, this E2EDM's prediction is incorrect, and the actual cause of the error (*e.g.*, the vehicle far from us on the right) becomes less noticeable as it is overshadowed by the many highlighted elements. Consequently, humans are unable to recognize the precise cause responsible for the wrong prediction, as human perception tends to believe what they want to believe. Therefore, complex explanations have the potential to deceive humans.

To address these problems, inspired by the Refined Objectification Branch (ROB) proposed in [29], we innovatively proposed two indicators that determine the explainability of E2EDMs: the objectification degree and the simplification degree. Although previous studies [21], [22], [23], [24], [25] and we aim to improve the explainability, they neglect to improve the simplification degree thus generating over-complex and deceivable explanations. On the other hand, we innovatively introduce the SOB structure, which not only improves the objectification degree of the explanations but also the simplification degree, as shown in Fig. 1 (iii), which results in simpler explanations. Furthermore, the integration of SOB into E2EDMs not only preserves prediction accuracy but actually enhances it. Lastly, SOB can be integrated into any type of existing E2EDMs without specific structural requirements, thus ensuring the high flexibility of our proposal.

The contributions of this paper are:

- We proposed the SOB structure, which can be integrated into any existing E2EDMs. The SOB enhances the objectification and simplification degree of explanations.
- We performed experiments to evaluate the explainability of E2EDMs. Based on the results, we demonstrated that SOB could improve the explainability of E2EDMs.

II. Related Work

In explainable AI (XAI), understanding the predictions of machine learning models has a general process. First, there is a target *model* that needs to be explained [30], [31], [32], [33]; then, we select an *explanation method* [12], [34], [35], [36]; finally, we use this explanation method to obtain *explanations* [37], [38], [39], [40]. Therefore, we introduce the previous research about 3 topics: The driving models that try to improve the explainability, explanation methods, and the properties of the explanations.

A. The Driving Models That Try to Improve the Explainability

1) The Object-Based Explainability Enhancement Approach: As introduced in the previous study of cognitive science [25], the basis of human attention is the object. Therefore, the previous studies integrate an object detection module into the driving models to generate object-level explanations, thereby making the driving models more explainable.

Chen et al. [22] and Sauer et al. [23] introduced human-interpretable intermediate features, such as lane curvature, distance to neighboring lanes, and distance from front-located vehicles. They first trained a convolutional neural network to produce these features and then mapped these

features to the steering angle. Wang et al. [20] extracted object bounding boxes and used the object feature to predict driving actions. In their later work [24], they further extracted 3D object information, such as depth, rotation, and size.

However, they share the shortcomings of specific structural requirements for the object detection module, thereby restricting their flexibility. In this paper, we solve this problem by introducing SOB, which encourages the E2EDMs to focus on the object feature without an object detection module.

2) The Attention-Based Explainability Enhancement Approach: Previous studies [41], [42], [43] introduced E2EDMs that are integrated with an attention mechanism to enhance their explainability. However, none of the above studies evaluated the explainability to validate the effectiveness of their proposals. In our previous work [29], based on the attention mechanism, we proposed a refinement branch and performed human experiments to evaluate explainability.

A common limitation in prior research (object-based and attention-based) is the disregard for the simplicity of explanations, which is another important property of explainability. In this paper, we explicitly improve the simplification degree by proposing the SOB structure, which could also be integrated into any E2EDMs to ensure its flexibility.

B. The Explanation Methods

We first divide various explanation methods into two different groups, the global and local explanation methods. For each group, we then divide them into two subgroups, the active and passive explanation methods.

1) Global Explanation Methods: In this category, we present explanation methods that aid in comprehending the whole decision-making process of a target model.

Global-Active Explanation Method: Previous studies [44], [45] developed a prototype classifier by adding a prototype layer. The network makes predictions based on the similarity between inputs and the learned prototypes, then the network provides prototypes as explanations.

Global-Passive Explanation Method: The primary global-passive explanation method is mimic learning, where a deep model is used as a teacher, and an interpretable shallow model is used as a student. The overall process can be regarded as a distillation process from the teacher to the student, where the interpretable student model provides a global view of the deep teacher model [46], [47], [48].

2) Local Explanation Methods: In this category, we present explanation methods that aid in comprehending the individual instance level by analyzing specific decisions.

Local-Active Explanation Methods: Local-active explanation methods involve the attention mechanism, which is used to explain specific predictions by identifying the essential features through attention weights [49], [50], [51], [52].

Local-Passive Explanation Methods: This category is the most commonly used for explaining deep learning models. This method can be further categorized into three types: gradient-based [53], [54], occlusion-based [55], and local approximation methods [56].

Grad-CAM [54] computes a saliency map with respect to a particular class on the last convolutional layer and can be used to explain any convolution-based models.

Zeiler and Fergus [55] proposed an occlusion-based method, where a gray patch is overlaid on the image, and the prediction changes are considered as the importance of the covered area.

Ribeiro et al. proposed LIME [56] to explain any model by local approximation. This method approximates the target model with an interpretable model, such as logistic regression, to provide explanations for individual predictions.

Compared to global explanation methods that aim to explain the whole model, the local explanation methods generate explanations for each decision, thus their explanations could be easily understood and evaluated. Therefore, we use local explanation methods to generate attribution-based explanations for each prediction. For E2EDMs with attention mechanisms (our proposal and baselines), we use the local-active explanation method [50], for E2EDMs without attention mechanisms, we use the local-passive explanation method [54].

C. Properties of the Explanations

We divide the properties of the explanations into two categories: those describing the relationship between the explanation and the target model, and those describing the relationship between the explanations and humans.

The properties that describe the relationship between the explanations and the target model

Fidelity [37] (correctness [38]): Yang et al. [37] define fidelity as the degree to which the explanation accurately represents the target model.

Completeness [39], [40], [57]: Cui et al. [39] define completeness as the degree to providing a complete explanation for the model's predictions.

The properties that describe the relationship between the explanations and the human

Complexity [26], [27], [28]: Kulesza et al. [26] define complexity that specifies an aspect of the explanations understanding process. The simpler the information from the explainer, the easier it is to understand for the explainee.

Persuasibility (Correlation [39], Interpretability [57]): Yang et al. [37] define persuasibility as the comprehensibility of an explanation. Despite being given different names, whether explanations can be understood by humans has always been a focal point of XAI research.

In simple tasks, such as object detection, where the human-labeled truth is consistent across various user groups, the persuasibility of an explanation can be objectively assessed by using annotation-based evaluation methods, such as bounding boxes and semantic segmentation [54]. However, in complex tasks, using human annotations to evaluate persuasibility may not be appropriate since relevant annotations may differ across various user groups. As a result, human-dependent experimental evaluation is a common method for evaluating the persuasibility of explanations in such tasks [19], [58] [59].

Compared to simple tasks, in complex driving tasks, how to improve the persuasibility of the explanation is more difficult

since it may vary across various user groups. Therefore, we propose two indicators that are valued in the human understanding of driving tasks, which are the simplification and objectification degrees. To explicitly enable E2EDMs to generate persuasive explanations that have high simplification and objectification degrees, we propose the SOB structure.

III. The Proposed Method

In this section, we first introduce the design concept of the SOB structure, then introduce its architecture, and finally the implementation details for SOB-integrated E2EDMs.

A. The Design Concept of SOB Structure

For the objectification degree, as introduced in the previous study of cognitive science [25], the basis of human attention is the object. Specifically, humans prefer object-level explanations in driving tasks [19]. For example, when facing a need-to-brake situation, humans prefer explanations to be a clear driving-related object, such as a "vehicle" instead of an undefinable "area". Therefore, we believe generating explanations that have a high objectification degree could make the E2EDMs more explainable.

For the simplification degree, to calculate the simplification degree of the explanations, we propose a method to produce the importance of objects from the perspective of E2EDM. Based on many previous studies [39], [40], [57], simplicity is vital for explainability. Therefore, we could make an explanation persuasive by making it simple, *i.e.*, the important objects that lead to the driving actions could be easily identified. In an ideal situation, all objects could be clearly distinguished into important objects and unimportant objects. If we constrain the importance of objects within the range of 0 to 1, where 0 represents unimportant objects and 1 represents important objects, the most simple explanation would be when the importance of objects is either 0 or 1. More specifically, when half of the objects are 0 and the other half are 1, the variance of all objects' importance would be maximal.

In a practical situation, we could consider the distribution of all objects' importance as a bi-modal distribution, which consists of two distributions: the important objects distribution and the unimportant objects distribution. The simple explanation has a clear separation of these two distributions, *i.e.*, the between-class variance is maximal. As introduced in [60], the between-class variance σ_B^2 could be calculated as

$$\sigma_B^2 = \sigma_T^2 - \sigma_W^2, \tag{1}$$

where σ_W^2 and σ_T^2 are the within-class variance and total variance. By drawing on the experience of this idea, we assume the within-class variance σ_W^2 is nearly constant, then we could separate important objects distribution from the unimportant objects distribution by maximizing σ_T^2.

Therefore, we define the standard deviation of the objects' importance as the simplification degree of an explanation, which is a simple and differentiable function for trainable E2EDMs. Generating explanations with a high simplification degree could make the explanations more understandable, thereby making the E2EDMs more explainable. The simplification degree and distribution of objects' importance will be thoroughly introduced in the section V-B.2.

The SOB structure consists of two branches: an objectification branch and a simplification branch. These two branches are designed to make the refined feature (as shown in Fig. 2) more object-centric and simplified. Since we apply the attention-based explanation method to generate explanations for SOB-integrated E2EDMs, the refined feature is not only used to make predictions about driving actions but also to generate explanations. In addition, the attention mechanism is applied to the last feature layer of the network, which ensures the attention-weighted feature is directly used for the final driving classification. Therefore, the generated explanations are faithful to the predictions about driving actions. Each branch could calculate a loss to represent to which extent the E2EDM could generate explanations that have a corresponding degree, e.g., the objectification branch calculates objectification loss, which represents to which extent the E2EDM could generate explanations that have high objectification degrees. By integrating the objectification and simplification loss from two branches into the loss function, the E2EDM's structure could have a specific connection with explainability.

B. The Architecture of SOB

1) The Objectification Branch: As shown in Fig. 2, we employ the modern semantic segmentation structure, *i.e.*, the fully convolutional network (FCN) [61] as the objectification branch to predict the area of objects (vehicles, pedestrian, lanes, traffic lights, *etc.*). We integrate the FCN with the E2EDMs' backbone, the FCN combines feature maps with different sizes from backbone [62] to predict object areas. For the ground-truth mask of object areas, based on the human annotation results, we assign pixel values of 1 to the object area and pixel values of 0 to other areas, we define the ground-truth mask as $O_{all}^{224\times224} \in \{0, 1\}$. By comparing the predicted object areas with the ground-truth object mask, we calculate the objectification loss as

$$\mathcal{L}_O = DCS(O_{all}, \hat{O}_{all}), \tag{2}$$

which will be used in the loss function. The DCS is the dice loss [63]. As shown in Fig. 2, O_{all} and $\hat{O}_{all}$ are the ground truth mask and prediction results of all objects' areas, respectively. Smaller $\mathcal{L}_O$ indicate the E2EDM could generate explanations with higher objectification degrees.

2) The Simplification Branch: The simplification branch is designed by extending the refinement branch in [29]. In this paper, the simplification branch computes the simplification loss, thus the E2EDMs could generate explanations with high simplification degrees by minimizing the simplification loss.

First, we introduce the refinement branch. As shown in Fig 2, the E2EDMs take an image (or images) as input, and after passing them through the backbone, we get a backbone feature of size C × H × W (C: channel, H: height, W: width). Then, we apply channel-wise and spatial-wise attention mechanisms to the backbone feature, resulting in an attention mask of the same size [52]. Finally, we element-wisely multiply the

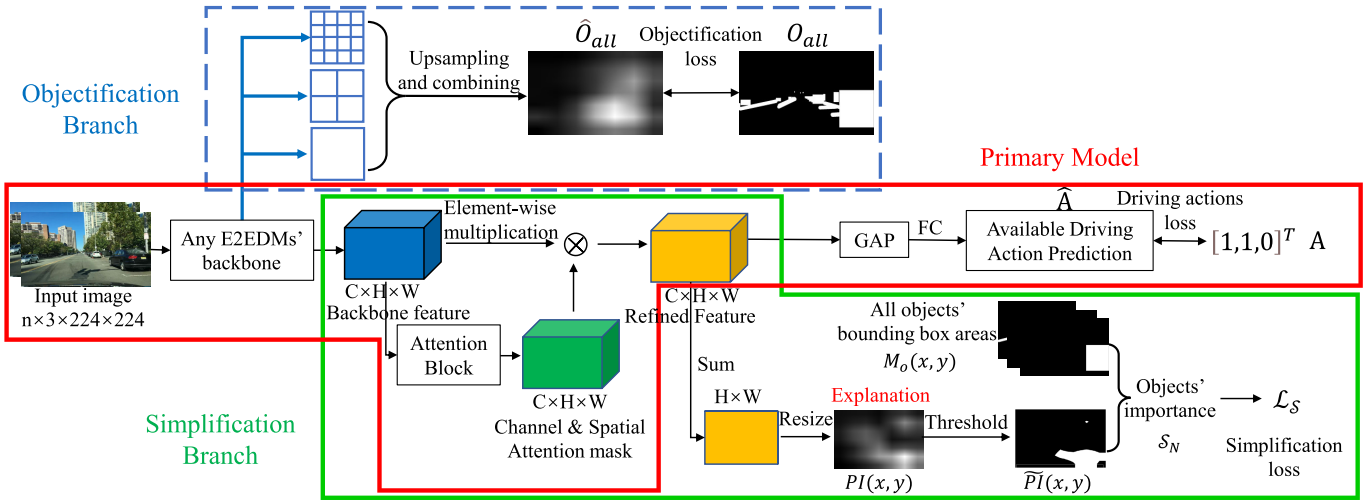

Fig. 2. The architecture of a SOB-integrated E2EDM. This architecture consists of three parts: the primary model, the objectification branch, and the simplification branch. **The primary model** aims to predict driving actions from multiple consecutive driving scene images, using a pretrained backbone to extract features and an attention mechanism (CBAM [52]) to refine features. These refined features directly contribute to action prediction and serve as explanations, ensuring faithful explanations for the predictions. **The objectification branch** predicts the locations of driving-related objects in the scenes, leveraging intermediate layer features and FCN [61], the prediction loss is introduced in Eq. (2). Finally, **the simplification branch** aims to make explanations easier to understand by reducing explanation complexity, measured by the standard deviation in objects' importance scores, which is introduced in Eq. (7).

attention mask with the backbone feature to obtain the refined feature,. The refined feature is used to predict driving actions and calculate the simplification degree.

Next, we introduce the additional structures to make the refinement branch upgrade to the simplification branch. First, we sum all C channels of the refined features, which leads to an H × W feature map. We then resize this H × W feature map to the size of the original images to gain the pixel-level importance map. We denote the pixel-level importance map as PI, and we identify its important area $\widetilde{PI}$ as:

$$\widetilde{PI}(x,y) = \begin{cases} 1 & \text{if } PI(x,y) \geq T_i \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $PI(x,y)$ represents the pixel value at position (x,y) in the pixel-level importance map, T_i is the threshold value as 0.5 to determine the important and the unimportant areas.

We define the area inside each object's bounding box as:

$$M_o(x,y) = \begin{cases} 1 & \text{if } (x,y) \in B_{box}(o) \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where $B_{box}(o)$ is the bounding box of each object, the pixel value inside the object's bounding box is 1, and the pixel value outside the bounding box is 0.

Next, we calculate the intersection over union (IOU) between $\widetilde{PI}$ and the bounding box area of each object as each object's importance:

$$I(o) = \frac{\widetilde{PI}(x,y) \cap M_o(x,y)}{\widetilde{PI}(x,y) \cup M_o(x,y)}, \tag{5}$$

where $I(o)$ represents the importance score of an object.

From a human perspective, most objects are unimportant in driving scenarios, if we directly maximize the standard deviation of all objects' importance, the most ideal situation would be 50% of all objects are unimportant and 50% of all objects all important, which is not realistic in most driving scenarios. Therefore, given the entire set of objects, we identify a subset containing the least important objects. By removing these unimportant objects, we aim to maximize the variance within the remaining subset (potentially important objects set), allowing us to further distinguish and separate the objects likely to be important. Consequently, approximately half of the potentially important objects set is deemed unimportant, while the other half is considered important. This approach ensures that a majority of objects are of low importance, while only a minority are considered highly important.

We denote $\mathbb{U}$ to be the set of importance scores for all objects, the $\mathbb{S}_M$ to be the set of importance scores for M potentially important objects in a driving environment as

$$\mathbb{S}_M = \{I(o_1), I(o_2), \ldots, I(o_M)\}, \tag{6}$$

e.g., when $M = 10$, it signifies that $\mathbb{S}_M$ encompasses the importance scores of top 10 important objects. $I(o_i)$ is the importance of the object ranked at i-th.

The M is set to the 70% of all objects in a driving environment. Then, we calculate the standard deviation of the potentially important objects $\mathbb{S}_M$ as the simplification loss

$$\mathcal{L}_S = -\sigma(\mathbb{S}_M), \tag{7}$$

which will be used in the loss function. σ is a function to calculate the standard deviation. By minimizing this loss, the object importance in the potentially important object set is demanded to be more dispersed, *i.e.*, important objects to be more important (importance close to 1), and less important objects to be less important (importance close to 0), thereby finding the true important objects from the potentially important object. Smaller $\mathcal{L}_S$ indicates the E2EDM could generate explanations that have higher simplification degrees.

C. Implementation Details

In this paper, our E2EDMs take two consecutive images as input (input size $2 \times 3 \times 224 \times 224$) and use pretrained backbones. We train the E2EDMs for 50 epochs with a multi-task loss function, which combines three components: driving action loss, objectification loss, and simplification loss. The Adam optimizer is utilized with a weight decay of 1×10^{-4} and an initial learning rate of 0.001.

The multi-task loss function is formulated as

$$\mathcal{L} = \lambda_A \mathcal{L}_A + \lambda_O \mathcal{L}_O + \lambda_S \mathcal{L}_S, \tag{8}$$

where $\mathcal{L}_A = BCE(A, \hat{A})$, the A and $\hat{A}$ denote the ground truth label and prediction result of driving actions, respectively. BCE is the binary cross entropy loss. λ_A, λ_O, and λ_S are hyperparameters that control the relative importance of driving action loss, objectification loss, and simplification loss, in this paper, they are set to 1, 1, and 0.02, respectively.

IV. Experiment

A. The BDD-3AA Dataset

Previous driving datasets [6], [8] primarily focused on designating the driver's chosen action as the ground truth for a driving scenario, suggesting that only that specific action was correct. However, drivers tend to select driving actions randomly from several correct options. As a result, these previous driving datasets carried the risk of training E2EDMs with an incomplete grasp of the full driving scenario, making them unsuitable for comprehensive evaluations of explanations.

To address this concern, we utilized the BDD-3AA (3 Available Actions) [19] dataset for training E2EDMs. Based on the environment information such as surrounding vehicles, pedestrians, lanes, and traffic lights, each driving scenario in the BDD-3AA dataset was annotated with the availability of three distinct driving actions: acceleration, steering left, and steering right. Thus, we treated the driving task as a multi-label classification problem. Among various driving tasks, classification tasks offer convenient methods to assess the persuasibility of explanations generated by E2EDMs. Consequently, such classification tasks stand out as optimal choices for evaluating explanations.

The BDD-3AA dataset comprises 500 video clips. When presented with successive images capturing the driving surroundings, the objective of the E2EDMs is to determine the availabilities for three distinct driving actions: acceleration, steering left, and steering right. As shown in Fig. 3, the ground truth for this typical scene is $A = [1, 1, 0]^T$, 1 indicates the corresponding driving action is available and 0 indicates unavailable, thus A indicating that acceleration and steering left actions are available while the steering right action is not.

To evaluate the prediction accuracy of our E2EDMs, due to the imbalance of driving actions in the dataset, *i.e.*, most acceleration actions are available, while most steering left and right actions are not available. Specifically, among 500 driving scenes, the acceleration actions of 450 scenes are available, the steering left actions of 175 scenes are available, and the steering right actions of 205 scenes are available. We utilized the macro F1 score to evaluate prediction accuracy, which involved computing the average F1 score of the three actions (acceleration, steering left, and steering right).

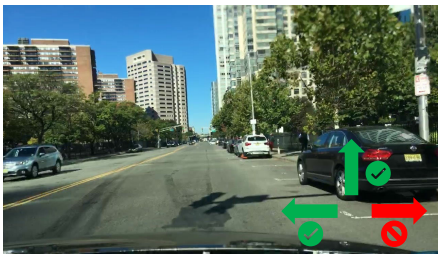

Fig. 3. Typical scene in the BDD-3AA dataset. As shown in the above image, there is a vehicle on the right, thus the steering right action is not available; there are vacant spaces in the front and left, thus the acceleration and steering left actions are available. In the bottom right of this image, the red arrow indicates the corresponding driving action is unavailable, the green arrow indicates the corresponding driving action is available.

$$Macro\ F_1 = \frac{F_1(\hat{A}_a, A_a) + F_1(\hat{A}_l, A_l) + F_1(\hat{A}_r, A_r)}{3}, \tag{9}$$

where A_a, A_l, A_r are the acceleration, steering left, and steering right actions.

B. The Experimental Persuasibility Evaluation Method

An experimental method to evaluate the persuasibility of explanations is proposed in [19]. We gathered 5 participants who possess driver's licenses. Each explanation is evaluated by at least three participants, we calculate the average value as the final score. These experimental methods consist of two methods: the driving action reproduction experiment and the heatmap judgment experiment.

1) The Driving Action Reproduction Experiment: This experiment determines whether explanations can correctly highlight driving-related features. We only show the most important part of an image to participants according to the explanations, if the participants can make the same annotation results based on this partially shown image as they would with a complete image, it means the explanations can correctly highlight driving-related features. We utilize the macro F1 score to measure the similarity between the annotation results of partially shown images and complete images. A higher score indicates more persuasive explanations.

2) The Heatmap Satisfaction Experiment: We assess the participants' satisfaction level with the explanations. Participants rate the heatmap (as shown in Fig. 1) from 1 to 5, with 1 being low persuasibility and 5 being high persuasibility.

C. The Objectification and Simplification (OAS) Explanation Evaluation Method

This method objectively evaluates the explanations generated by E2EDMs without using humans as participants. The evaluation is divided into two indicators: the objectification degree and the simplification degree, thus the name of the evaluation method is OAS (objectification and simplification). As we introduced before, the objectification degree represents the extent to which driving-related objects are utilized; the simplification degree represents the simplicity of the explanation, i.e., the dispersity of the objects' importance. Given that

the human recognition system relies on objects, and simple explanations are more comprehensible to humans, objectification and simplification degrees determine the persuasibility of the explanation, which is closely related to the explainability of the E2EDMs. Therefore, we could provide a thorough analysis of how the SOB structure improves the explainability of E2EDMs, offering robust validation of our proposal. The evaluation is divided into two indicators: the objectification degree and the simplification degree.

1) The Objectification Degree:

$$OD = \frac{\sum_{p \in O_{all}} L(p)}{\sum_{p} L(p)}, \tag{10}$$

where OD is short for Objectification Degree, $L(p)$ represents the luminance of a pixel in the explanations, similar to the pixel value of PI in Fig. 2, which also is the importance score assigned to a pixel, $\sum_{p \in O_{all}} L(p)$ represents the summation of all pixels' importance scores inside the ground-truth object mask, $\sum_{p} L(p)$ represents the summation of all pixels' importance scores in the explanations.

2) The Simplification Degree:

$$SD = \sigma(\mathbb{U}) \times 100, \tag{11}$$

where the SD is short for Simplification Degree, the calculation of the set of objects' importance is the same as Eqs. (3) $\sim$ (6), the differences are that the PI in Eq. (3) is replaced by all kinds of generated attribution-based explanations, which also contain the importance of each pixel, and instead of calculating the standard deviation of potentially important objects $\mathbb{S}_M$, we calculate the importance of all objects $\mathbb{U}$.

V. Experimental Results and Discussion

In this section, we present experimental results to demonstrate the contribution we mentioned in the introduction.

A. Validation of the SOB Could Better Improve the Explainability and Prediction Accuracy of E2EDMs

We evaluate the explainability and prediction accuracy of 4 E2EDMs with a ResNet-18 backbone. These 4 E2EDMs are:

- Vanilla: the baseline E2EDM without any branch.
- RB: the E2EDM with only the refinement branch, which is equal to the simplification branch before summing the refined features, similar to previous studies [41], [42], the RB is based on attention mechanism.
- ROB: the E2EDM with both the objectification branch and the refinement branch [29].
- SOB: the E2EDM with both the objectification and simplification branches.

1) The Experimental Explainability Evaluation Results: Throughout the experiments, participants were kept unaware of the prediction results generated by the E2EDMs. For each driving scene, three driving actions must be considered during the experiments. 1. Participants' annotation results about their judgment on the driving actions. 2. The ground truth for driving actions used for training the E2EDMs. 3. E2EDMs' prediction results for driving actions.

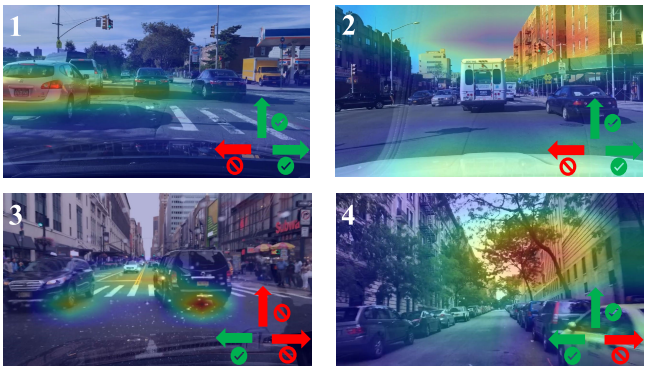

Fig. 4. These 4 images represent E2EDM's prediction results for each driving scene (in the bottom right of each image, the red arrow indicates the corresponding driving action is unavailable, the green arrow indicates the corresponding driving action is available.). The heatmap is the corresponding explanation for these predictions. In each image, the upper right number relates to a specific situation in the Table. I, e.g., for the upper right image 1, it is situation 1 in Table. I, the prediction is correct and the explanation is persuasive, thus this situation indicates high explainability.

TABLE I

Determine Explainability Based on the Relationship Between the Correctness of Prediction Results and the Persuasibility of the Explanations. As Shown in Fig. 4, We Further Illustrate Each Situation in the 2 × 2 Table With Examples, i.e., the E2EDMs' Prediction Results and the Explanations for These Predictions

	Persuasive explanation	Non-persuasive explanation
Correct prediction	1.High explainability	2.Low explainability
Wrong prediction	3.Low explainability	4.High explainability

Since even for the same driving scene, humans may have different opinions on driving actions, we must ensure participants' annotated driving actions match the ground truth for E2EDMs, thus we could ensure the participants have the qualifications to evaluate the explanations. Moreover, whether the E2EDMs could correctly predict driving actions will greatly influence the evaluation of the E2EDMs' explainability.

We divide the relationship between the correctness of prediction results and the persuasibility of the explanations into 4 situations. As shown in Table. I and Fig. 4, we introduce each situation, and for each situation, we show an example of E2EDM's prediction results and explanations.

1 The E2EDM's prediction is correct, and the explanations are persuasive. Thus the explanations could convince people that the computational method of E2EDM is correct, which indicates high explainability.

2 The E2EDM's prediction is correct, and the explanations are non-persuasive, i.e., people are unable to comprehend or trust that the E2EDM's computational method is correct. It indicates low explainability.

3 The E2EDM's prediction is wrong, and the explanations are persuasive, i.e., the explanations mislead people into believing in the E2EDM's computational method, which indicates low explainability.

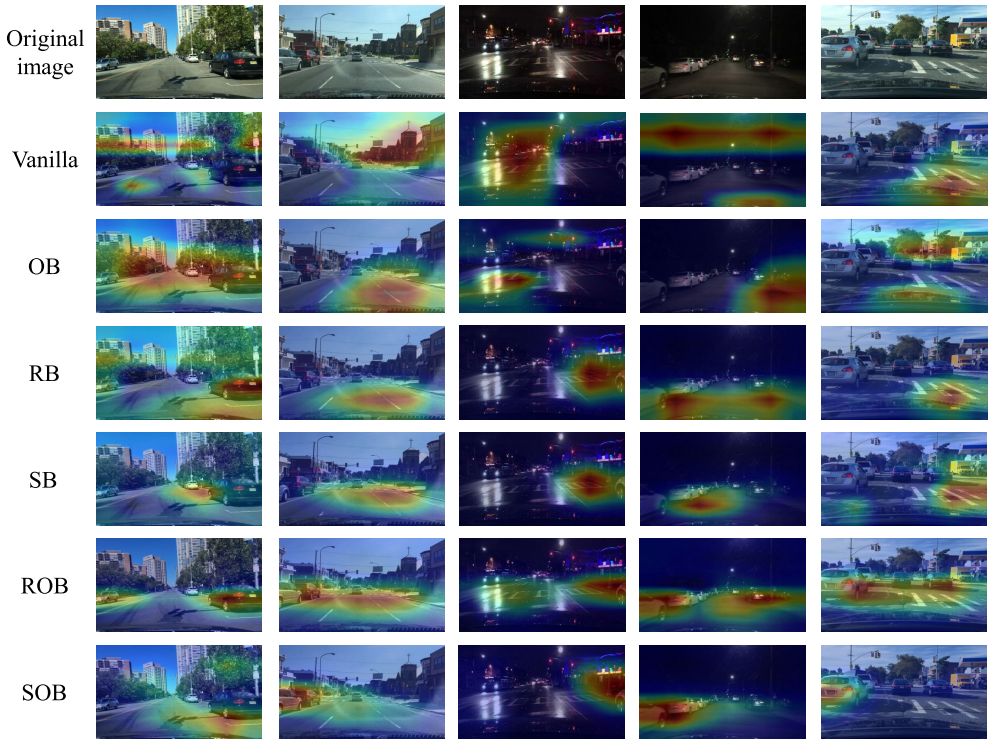

Fig. 5. The explanations generated from 6 E2EDMs in the ablation study. Besides ROB, SOB, RB, and Vanilla, the other 2 E2EDMs are OB (the E2EDM with only the objectification branch), and SB (the E2EDM with only the simplification branch). Based on previous research [19], to generate explanations for the predictions of the E2EDMs, we need to visualize the high-level features that are used to predict the driving action. Therefore, reasonable explanation results for OB and Vanilla should come from the feature maps generated by the last convolutional layer, while for E2EDMs with attention mechanisms, explanation results should come from the refined features in Fig. 2. Therefore, for Vanilla and OB, we used Grad-CAM to generate explanations, and for SOB, ROB, SB, and RB, we used an attention explanation method to generate explanations.

4 The E2EDM's prediction is wrong, and the explanations are non-persuasive, i.e., people recognize that the computational method is incorrect, indicating high explainability.

In summary, to ensure high explainability of E2EDMs, the persuasibility of explanations must align with the correctness of the predictions, i.e., if the prediction is correct, then the explanations should be persuasive; conversely, if the prediction is wrong, then the explanation should be non-persuasive.

Therefore, to assess the explainability of the E2EDMs, we evaluate the explanations under two different scenarios. In the first scenario, the E2EDM's prediction is correct, the explanations generated by E2EDMs are expected to be persuasive to the participants. In the second scenario, the E2EDM's prediction is wrong, and the explanations generated by E2EDMs are not likely to be persuasive for the participants. Since participants were kept unaware of the prediction results generated by the E2EDMs, participants' assessment of the persuasibility of the explanations is solely based on the explanations themselves. This allows us to evaluate the extent to which the explanations deceive human judgment.

To quantify this deceptive aspect, we introduce the *deceptive level*. This level is calculated by comparing the satisfaction scores of the generated explanations between the two aforementioned conditions (1st and 2nd scenarios). The deceptive level captures the difference in how convincing the explanations are perceived by the participants, revealing the degree to which the explanations manage to mislead human perception, the deceptive level is calculated as

$$Deceptive\ level = \frac{HS_{wrong}}{HS_{correct}}, \tag{12}$$

where $HS_{correct}$ denotes the heatmap satisfaction of the explanations when the E2EDMs made the correct predictions. HS_{wrong} denotes the heatmap satisfaction of the explanations when the E2EDMs made the wrong predictions.

When the E2EDM's prediction is wrong, the corresponding explanations should be non-persuasive, i.e., the heatmap satisfaction should be low; when the E2EDM's prediction is correct, the corresponding explanations should be persuasive, i.e., the heatmap satisfaction should be high. Therefore, a high deceptive level indicates that no matter whether the E2EDM makes the correct predictions, the generated explanations always mislead people to believe the E2EDM.

As shown in the first and second rows of Table. II, when the E2EDM's prediction is correct, the SOB generates more persuasive explanations than ROB, RB, and Vanilla. On the other hand, as shown in the third row of Table. II, the deceptive level of ROB, RB, and Vanilla is higher than SOB. This demonstrates when these E2EDMs could not handle the driving environment, they still made the participants believe that they could. We believe the reason is that previous E2EDMs tend to generate overcomplicated explanations that highlight too many objects in the images (as shown in Fig. 5), thus misleading the participants to believe that the ROB has the correct driving methods. Whereas the SOB focuses on fewer

TABLE II
THE EXPERIMENTAL EVALUATION RESULTS FOR THE EXPLANATIONS

Model	SOB	ROB	RB	Vanilla
Heatmap Satisfaction	**3.69**	3.67	3.35	2.23
Driving action reproduction score	**82.1%**	79.8%	77.9%	68.5%
Deceptive level	**0.86**	0.94	0.95	1.17

TABLE III
THE OAS EVALUATION RESULTS FOR EXPLANATIONS FROM 3 BASELINE E2EDMs AND THEIR CORRESPONDING SOB-INTEGRATED E2EDMs

Model		OAS	
		OD	SD
CBAM [52]	RB	0.28	3.86
	SOB	0.37	**6.25**
CCnet [51]	RB-CCnet	0.31	4.57
	SOB-CCnet	**0.40**	6.22
ABN [43]	RB-ABN	0.20	2.44
	SOB-ABN	0.32	5.03

objects, thus the participants can grasp the exact cause behind the wrong predictions and can not be misled.

2) Validation of the Effectiveness of SOB Structure Based on the OAS Explanations Evaluation Results: To further demonstrate the effectiveness of SOB structures, besides our previous study [29], we integrate SOB into other baselines [43], [51].

In RB, the refinement branch is designed based on CBAM [52], an attention-based structure that could generate an attention mask based on backbone features. The attention mask is applied to backbone features to produce the refined features for the latter calculation in the simplification branch, i.e., the main proposal in this paper. To test our proposal in other baselines, we replace the current refinement branch (CBAM) with other attention-based structures to generate refined features.

Mori et al. [43] proposed an attention branch network (ABN), and Woo et al. [51] proposed criss-cross attention (CCnet). Similar to CBAM [52], these two structures are both attention-based structures, they could also generate attention masks based on the backbone features, and thus they could both serve as refinement branches. We integrate these two structures (ABN and CCnet) into Vanilla, respectively. We denote these two models as RB-ABN and RB-CCnet, which as baseline models from other studies.

Similar to how we integrate SOB structures to the RB to get SOB, we integrate the SOB structure to RB-ABN and RB-CCnet to obtain SOB-ABN and SOB-CCnet. We use the OAS explanation evaluation method to evaluate the explanations generated by these 3 pairs of E2EDMs, each pair contains a baseline E2EDM and its SOB-integrated E2EDM. As shown in Table. III, no matter for which indicator (objectification degree or simplification degree), the SOB-integrated E2EDMs all outperform their corresponding baseline E2EDMs. Therefore, SOB could help the E2EDMs generate more persuasive explanations, i.e., help the E2EDMs become more explainable.

The reason behind the advantage of SOB in persuasibility is that it considered the simplification degree of the explanations. Along with the objectification degree, we believe these two indicators together determine the persuasibility of the explanations generated by the E2EDMs. By integrating the simplification loss and objectification loss into the loss function, the SOB structures can train E2EDMs' ability to generate explanations that have high objectification and simplification degrees, as shown in Fig. 5, the explanations from SOB tend to focus on the most important objects in the images, leading to improved explainability. Compared to the previous baseline (ROB), SOB utilizes more information during the training process. Specifically, ROB uses driving action labels and a mask representing the location information of all objects, while SOB uses driving action labels and multiple masks, each representing the location information of individual objects. By knowing the location information of each object, SOB gains an ability that ROB lacks: calculating the importance score of each object from the model's perspective. SOB can use this information to more explicitly direct the model's focus towards the most important objects and disregard those that are unimportant. Therefore, the stronger persuasibility of SOB's explanations may be attributed to its use of more label information in training and the model's effective structure (simplification branch) to utilize this additional label information. However, since the individual object location information could be easily acquired, the high performance of SOB does not rely on expensive annotation labels.

In this paper, our main contribution is the combination of two simple branches: the objectification branch is a basic semantic segmentation structure, while the simplification branch is based on the attention module. However, the combination of these two simple structures can significantly improve the explainability of the E2EDMs. In addition, explainability encompasses other factors, such as fidelity, therefore, we plan to design more advanced structures to comprehensively and further improve the E2EDM's explainability in future work.

3) Performance of the SOB in Other Autonomous Driving Datasets: We showed the performance of SOB-integrated E2EDMs on the BDD-3AA dataset. However, the driving scenes included in the BDD-3AA dataset are not comprehensive enough. Therefore, we test SOB (trained on the BDD-3AA dataset) on mainstream driving datasets. This allows us to assess whether our proposal can handle a more diverse range of driving scenarios, i.e., whether it can make correct driving predictions and have high explainability.

We test SOB on nuScenes [64], Waymo [65], and Tusimple [66]. These datasets are renowned in the autonomous driving field and differ from BDD-3AA in specific aspects: 1. nuScenes includes driving environments in Singapore. 2. Waymo offers higher-resolution driving images. 3. Tusimple focuses mainly on highway driving scenarios. As shown in Fig. 6, for various scenes from these datasets, the SOB makes accurate predictions across different environments and provides persuasive explanations.

4) Performance of the SOB in Difficult Driving Scenes: In the previous sections, we demonstrated the performance of our E2EDMs across many datasets in terms of prediction accuracy and high explainability. However, autonomous driving is a complex task involving complicated driving scenarios, one of which includes driving scenes with red traffic lights. As shown in Fig. 7, in such scenarios, the red light is critical information

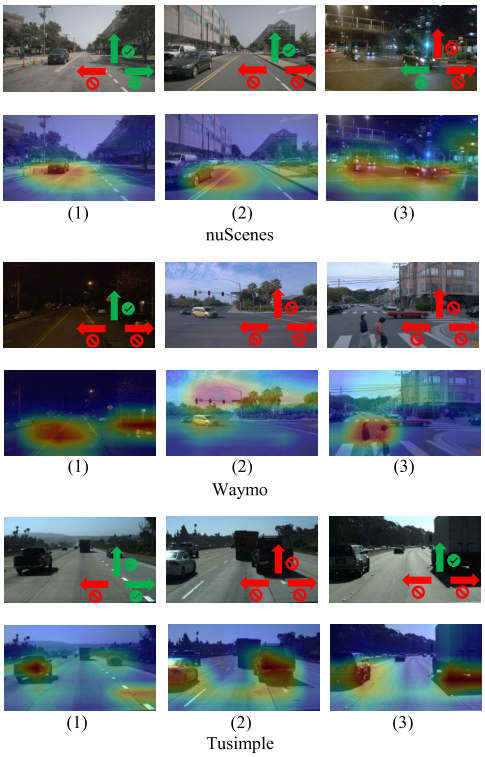

Fig. 6. There are SOB's predictions and explanations for 3 datasets. For each dataset, the first row is the prediction results, the second row is the corresponding explanations. We can see that not only the prediction results are correct but also the explanations. The SOB could make correct prediction results based on the right objects, such as vehicles; lanes, e.g., (2) in nuScenes, (1) in Waymo, (1) in Tusimple; pedestrians, e.g., (3) in Waymo; traffic light, e.g., (2) in Waymo.

Fig. 7. There are SOB's predictions and explanations for the red traffic light driving scenes. The left image is the prediction results, and the right one is the corresponding explanations. We can see that not only the prediction result is correct but the attention of the SOB is also focused on the red traffic light.

for determining driving actions. Given its small size compared to the entire image, it is difficult for E2EDMs to utilize the red light to make accurate predictions about driving actions.

Therefore, it is necessary to demonstrate the performance of our proposal in such difficult driving scenarios. As shown in Fig. 7, we can see the predictions made by our SOB for the availability of driving actions in Fig. 7, along with the explanations generated for these predictions. We can see that not only are the SOB's predictions accurate, but the SOB also correctly utilizes the red light in making these predictions.

5) The Improvement of SOB Structure on the Prediction Accuracy of E2EDMs: To assess the impact of the SOB structure on prediction accuracy, we integrated it with several widely-used backbones, including ResNet-18 [62], DenseNet [67], MobileNet [68], Inception [69], and ShuffleNet [70]. For each backbone, we train the SOB-integrated E2EDMs to compare them to the respective ROB-integrated E2EDMs, RB-integrated E2EDMs, and Vanilla E2EDMs. As shown in Table. IV, all SOB-integrated E2EDMs exhibit superior prediction accuracy. In summary, the multi-task training method designed to improve explainability could also help achieve higher prediction accuracy. However, due to the imbalance of driving actions in the dataset, i.e., most acceleration actions are available, while most steering left and right actions are not available, it will affect the prediction accuracy of the E2EDMs. We use the prediction accuracy of our main proposal, the SOB (ResNet-18 as backbone) as an example, the F1-scores for three distinct driving actions: acceleration, steering left, and steering right are 94.53%, 62.54%, 65.24%. We could see that since in most scenes, the ground truths of acceleration actions are available, predicting the availability of the acceleration actions is much easier. In the future, we plan to fix this problem by using up-sampling and down-sampling to make a more balanced dataset or use the focal loss [71] to train the E2EDMs to improve the prediction accuracy of steering left actions and steering right actions.

We believe the SOB structure could improve the prediction accuracy of any E2EDMs, in the future, we plan to use more diverse backbones, such as vision transformers, to further verify the effectiveness of our SOB structure.

To discuss the trade-off between explainability and prediction accuracy [72], we show 7 E2EDMs' prediction accuracy and simplification degree of the generated explanations in Fig. 8. The RB showed better explainability and worse prediction accuracy than Vanilla. However, the ROB and SOB both

TABLE IV
THE PREDICTION ACCURACY OF SOB, ROB, RB, AND VANILLA ON FIVE BACKBONES

Backbone	SOB	ROB	RB	Vanilla
ResNet-18 [62]	**74.10%**	72.87%	72.06%	72.53%
DenseNet-121 [67]	**73.92%**	72.56%	70.33%	72.23%
ShuffleNet-V2 [70]	**70.26%**	69.49%	68.20%	69.30%
MobileNet-V2 [68]	**74.78%**	74.32%	72.12%	72.49%
InceptionNet-V3 [69]	**75.78%**	73.33%	71.81%	72.23%

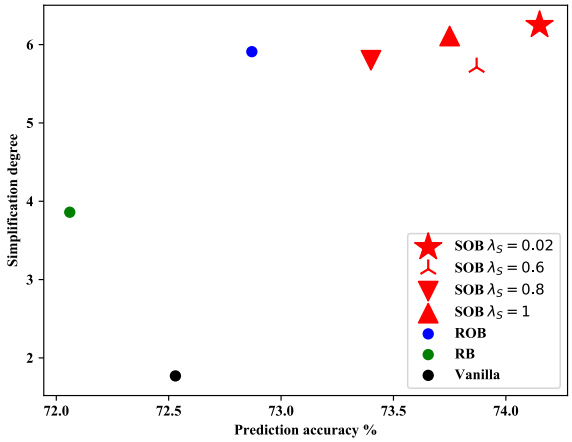

Fig. 8. To show whether there are trade-off phenomena, we show the prediction accuracy and simplification degree of the generated explanations about 7 E2EDMs which include the ROB, RB, Vanilla, and 4 SOBs. Since the main proposal of this paper is the simplification branch, we adjust λ_S, i.e., the hyperparameter that controls the relative importance of simplification loss to show whether it could lead to the trade-off phenomena. Note, the SOB where $\lambda_S = 0.02$ is the main proposal in this paper.

TABLE V
THE OAS EVALUATION RESULTS FOR EXPLANATIONS OF SOB AND SOB-L1

$\mathcal{L}_S$ calculation method	OAS	
	OD	SD
SOB	**0.37**	**6.25**
SOB-L1	0.36	5.65

have better explainability and prediction accuracy. Moreover, when the hyperparameter that controls the relative importance of simplification loss is adjusted, there is also no solid evidence to support the trade-off phenomenon. This observation may be due to the nature of the driving task in the BDD-3AA dataset. In this driving task, the prediction of the driving actions does not require all objects in the image. Meanwhile, the ROB and SOB excel at focusing on a small, crucial set of objects, thus leading to both higher explainability and prediction accuracy than the Vanilla, which lacks simplicity in its prediction methods. We suggest that the advantage of ROB and SOB may not hold in more complex driving tasks where all objects in the environment are crucial. In such cases, Vanilla's prediction methods of considering all objects might lead to more accurate predictions and less explainability than the SOB and ROB, i.e., the trade-off phenomenon.

B. Ablation Studies About Model Structures and Loss Function for Simplification Loss

1) Ablation Study About Loss Function for Simplification Loss: As shown in Eq. (7), the simplification loss is calculated by applying standard deviation to the potentially important objects $\mathbb{S}_M$. The simplification loss serves an important role in our SOB structures, it encourages the refined feature to become more simplified, thus the SOB-integrated E2EDMs could generate more persuasive explanations.

By minimizing the simplification loss calculated by standard deviation, the object importance in the potentially important object set is demanded to be more dispersed. However, The L1 norm is more widely used as a shrinkage method for feature selection [73], which could also be used to make the objects' importance become more dispersed. Therefore, we replace the standard deviation with the L1-norm to calculate the simplification loss, we denote this E2EDM as SOB-L1. We evaluate the explanations of SOB-L1 and the original SOB by the OAS explanation evaluation method. As shown in Table. V, no matter for which indicator (objectification degree or simplification degree), the SOB outperforms SOB-L1.

2) Ablation Study About Model Structures for Simplification Loss: When it comes to the evaluation of explanations, the *persuasibility* is too abstract and ambiguous for precise analysis. Thanks to the OAS explanations evaluation method that splits *persuasibility* into two explicit indicators, we can thoroughly discuss the impact of the SOB structure on explanations. In addition, to better analyze the effect of each branch, we use the OAS persuasibility evaluation method to evaluate the explanations generated by other E2EDMs as ablation studies: the E2EDM with only the objectification branch (OB), the E2EDM with only the simplification branch (SB).

TABLE VI
THE OBJECTIFICATION AND SIMPLIFICATION DEGREE OF 6 E2EDMs' EXPLANATIONS

Model		SOB	SB	ROB	RB	OB	Vanilla
OAS	OD	0.37	0.30	0.39	0.28	0.13	0.15
	SD	6.25	4.47	5.91	3.86	1.58	1.77

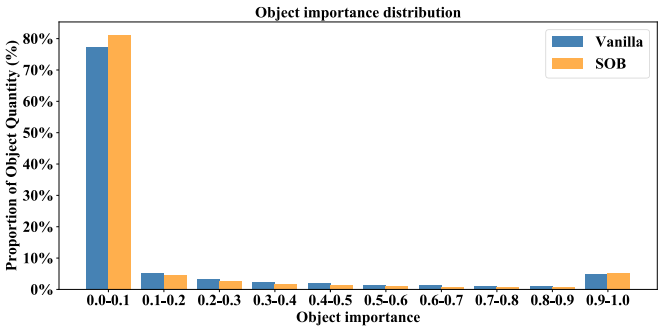

Fig. 9. The comparison of the object important distributions of SOB and Vanilla, showing our proposed branches could help the genuinely important objects stand out more prominently by reducing the number of moderately important objects and enlarging the number of unimportant objects.

First, for the objectification branch. We find that integrating the objectification branch on the E2EDMs does not always lead to an improvement in the objectification degree. As shown in the first row of Table. VI, when we integrate the objectification branch on the Vanilla to make it OB, the objectification degree does not rise. However, when we integrate the objectification branch on the SB to make it SOB, the objectification degree rises.

We speculate on the reasons behind these results. For the OB, the objectification branch predicts object areas by the FCN structure. The FCN structure uses feature maps from backbones to predict object areas, we believe the shallow layers from the backbone have the most ability to predict the object regions. However, when generating explanations for OB, we only visualize the last layer of the backbone. Therefore, a single objectification branch could not improve the objectification degree of explanations.

However, for E2EDMs equipped with a simplification branch, the explanations of the E2EDM are generated based on the attention mask. By integrating the simplification branch on the OB, the attention mechanism module could repair and enhance the ability to extract object elements as explanations. As a result, SOB has a better objectification degree.

The impact of the simplification branch on the simplification degree is consistent. As shown in the second row of Table. VI, when we add a simplification branch to the Vanilla, Vanilla becomes SB, and the simplification degree rises. Similarly, when we add a simplification branch to the OB, OB becomes SOB, and the simplification degree also rises.

To help readers better understand the impact of our proposed structure to the simplification degree, as shown in Fig. 9, we display the distribution of object importance from the perspective of SOB and Vanilla. For each E2EDM, based on the generated explanations, we calculate all objects' importance of each image from this E2EDM's perspective. After we combine the normalized objects' importance of all images

in the dataset and show them in Fig. 9, we could see the SOB has more unimportant objects than Vanilla (object importance $0.0 \sim 0.1$), meanwhile, the SOB also has less moderately important objects than Vanilla (object importance $0.1 \sim 0.9$), thus the genuinely important objects (object importance $0.9 \sim 1.0$) from SOB's perspective are easier to be distinguished from all objects, *i.e.*, the SOB is more explainable.

For the impact of the objectification branch on the simplification degree. As shown in the second row of Table. VI, when we add an objectification branch to the Vanilla, the simplification degree does not rise. However, when we add an objectification branch to the SB, the simplification degree rises. In summary, a single objectification branch could not improve the simplification degree, however, when it is combined with a simplification branch, it has a positive effect on the simplification degree.

Finally, we discuss the impact of the simplification branch on the objectification degree. As shown in the first row of Table. VI, after adding a simplification branch to the Vanilla, the objectification degree rises. After adding a simplification branch to the OB, the objectification degree rises. In summary, the simplification branch has a consistently positive impact on the objectification degree.

Note, that the simplification branch is an upgraded version of the refinement branch. As shown in the first and the second row of VI, both SB and SOB exhibit a higher degree of simplification compared to RB and ROB, while maintaining a similar degree of objectification to that of RB and ROB, indicating the simplification branch is a more powerful structure than the refinement branch. This is because the simplification branch explicitly improves E2EDMs' ability to make simplified explanations.

VI. Conclusion

In this paper, we proposed the SOB, which can be integrated into any existing E2EDMs to improve the E2EDM's explainability by making more simplified explanations, in addition, the SOB structure could also improve the E2EDM's prediction accuracy. In this paper, we focused on exploring the effectiveness of SOB structure in driving action classification tasks. For more complex driving tasks, such as predicting steering angles, due to the difficulty of conducting explanations evaluation experiments for such tasks, we leave this for future work.

We also proposed the OAS explanations evaluation method that does not rely on humans. This method enables us to quantitatively evaluate the E2EDM's explanations, which can help us better understand the impact of the different structures on the model's explainability. In the future, we plan to complete this method so it could replace human experiments to evaluate the persuasibility of the explanations.

At the end of this paper, we would like to advocate the core idea and the outlook for future work. In the process of machine learning development, at first, hand-craft feature extractors are used to extract features and perform predictions. Later, deep learning models with the autonomous search for the best features emerged and completely outperformed the previous methods. However, recently, many people have begun to try to design an interpretable model (*e.g.*, integrate object detection module to make it no longer pure end-to-end). This is going back to the old way and abandoning the successful deep-learning models, we believe that the correct logic is to determine a measure for human acceptance of the model's explainability (e.g., the objectification and simplification degree), thus the model could autonomously learn a more understandable prediction method and learn to present its prediction method in a more understandable way.

In this paper, we showed the potential that exists in this field. However, there is still a long way to go in this direction. Compared with optimizing the prediction accuracy of the model, optimizing the explainability of the model will be influenced by more factors. There is a multitude of endeavors that we must undertake to explore the impact of various factors on the explainability of models, including the type of task being addressed, the model structure, the loss function, and the choice of explanation methods. These investigations constitute the primary focus of our future research.

Acknowledgment

Chenkai Zhang would like to take this opportunity to thank the "Interdisciplinary Frontier Next-Generation Researcher Program of Tokai Higher Education and Research System."

References

[1] J. Lee and N. Moray, "Trust, control strategies and allocation of function in human–machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, Oct. 1992.

[2] 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*.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[4] J. Levinson et al., "Towards fully autonomous driving: Systems and algorithms," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 163–168.

[5] R. McAllister et al., "Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4745–4753.

[6] M. Bojarski et al., "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.

[7] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1988, pp. 1–9.

[8] 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. 3530–3538.

[9] A. Tampuu, T. Matiisen, M. Semikin, D. Fishman, and N. Muhammad, "A survey of end-to-end driving: Architectures and training methods," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1364–1384, Apr. 2022.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, Sep. 2019.

[11] 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.

[12] 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.

[13] M. Bojarski et al., "VisualBackProp: visualizing CNNs for autonomous driving," 2016, *arXiv:1611.05418*.

[14] 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.

[15] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "HiLM-D: Towards high-resolution understanding in multimodal large language models for autonomous driving," 2023, *arXiv:2309.05186*.

[16] H. Ben-Younes et al., "Driving behavior explanation with multi-level fusion," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108421.

[17] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," 2023, *arXiv:2310.01412*.

[18] B. Jin et al., "ADAPT: Action-aware driving caption transformer," 2023, *arXiv:2302.00673*.

[19] C. Zhang, D. Deguchi, Y. Okafuji, and H. Murase, "More persuasive explanation method for end-to-end driving models," *IEEE Access*, vol. 11, pp. 4270–4282, 2023.

[20] 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.

[21] 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. 9520–9529.

[22] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "DeepDriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2722–2730.

[23] A. Sauer, N. Savinov, and A. Geiger, "Conditional affordance learning for driving in urban environments," in *Proc. Conf. Robot Learn.*, 2018, pp. 237–252.

[24] D. Wang, C. Devin, Q.-Z. Cai, P. Krähenbühl, and T. Darrell, "Monocular plan view networks for autonomous driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 2876–2883.

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

[26] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proc. 20th Int. Conf. Intell. User Interfaces*, Mar. 2015, pp. 126–137.

[27] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[28] S. J. Read and A. Marcus-Newhall, "Explanatory coherence in social explanations: A parallel distributed processing account," *J. Personality Social Psychol.*, vol. 65, no. 3, pp. 429–447, 1993.

[29] C. Zhang, D. Deguchi, and H. Murase, "Refined objectification for improving end-to-end driving model explanation Persuasibility," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–6.

[30] V. Beaudouin et al., "Flexible and context-specific AI explainability: A multidisciplinary approach," 2020, *arXiv:2003.07703*.

[31] 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, doi: 10.1016/J.INFFUS.2019.12.012.

[32] A. Rosenfeld and A. Richardson, "Explainability in human-agent systems," *Auto. Agents Multi-Agent Syst.*, vol. 33, no. 6, pp. 673–705, Nov. 2019.

[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, 2018.

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

[35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[36] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

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

[38] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Trans. Interact. Intell. Syst.*, vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.

[39] X. Cui, J. M. Lee, and J. Hsieh, "An integrative 3C evaluation framework for explainable artificial intelligence," in *AI and Semantic Technologies for Intelligent Information Systems*. Cancún, Mexico: AIS eLibrary, 2019, pp. 1–10.

[40] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *Proc. IEEE Symp. Vis. Lang. Human Centric Comput.*, Sep. 2013, pp. 3–10.

[41] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[42] L. Cultrera, L. Seidenari, F. Becattini, P. Pala, and A. Del 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. 1389–1398.

[43] K. Mori, H. Fukui, T. Murase, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Visual explanation by attention branch network for end-to-end learning-based self-driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1577–1582.

[44] O. Li et al., "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 3530–3537.

[45] C. Chen et al., "This looks like that: Deep learning for interpretable image recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8930–8941.

[46] Q. Zhang et al., "Interpreting CNN knowledge via an explanatory graph," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 1–10.

[47] N. Frosst and G. Hinton, "Distilling a neural network into a soft decision tree," 2017, *arXiv:1711.09784*.

[48] 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. 6254–6263.

[49] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," 2019, *arXiv:1901.06706*.

[50] D. Mascharka, P. Tran, R. Soklaski, and A. Majumdar, "Transparency by design: Closing the gap between performance and interpretability in visual reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4942–4950.

[51] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[52] S. Woo et al., "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

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

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

[55] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Comput. Vis. ECCV 13th Eur. Conf.*, Zurich, Switzerland. Cham, Switzerland: Springer, Sep. 2014, pp. 818–833.

[56] 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*, 2016, pp. 1135–1144.

[57] L. H. Gilpin et al., "Explaining explanations: An approach to evaluating interpretability of machine learning," 2018, *arXiv:1806.00069*.

[58] 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.

[59] I. Lage et al., "An evaluation of the human-interpretability of explanation," 2019, *arXiv:1902.00006*.

[60] P. E. Hart, D. G. Stork, and R. O. Duda, *Pattern Classification*. Hoboken, NJ, USA: Wiley, 2000.

[61] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[62] 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.

[63] C. H. Sudre et al., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Third Int. Workshop, DLMIA 7th Int. Workshop ML-CDS Held Conjunct. (MICCAI)*, Montreal, QC, Canada. Cham, Switzerland: Springer, 2017, pp. 240–248.

[64] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[65] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2443–2451.

[66] *TuSimple: Tusimple Benchmark*. Accessed: Nov. 2019. [Online]. Available: https://github.com/TuSimple/tusimple-benchmark

[67] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[68] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[69] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[70] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[72] J. Huysmans, B. Baesens, and J. Vanthienen, "Using rule extraction to improve the comprehensibility of predictive models," Katholieke Universiteit Leuven, FETEW, Res. Rep. KBI 0612, 2006.

[73] T. Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.

Chenkai Zhang received the B.Eng. and B.A. degrees from 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 from 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 Post-Doctoral Fellow with Nagoya University in 2006. From 2008 to 2012, he was an Assistant Professor with the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor with the Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. He is currently working on object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as the detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.

Jialei Chen received the B.Eng. and M.Eng. degrees from Northeastern University, Shenyang, China, 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 semantic segmentation and image processing.

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 Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York, NY, USA. Since 2003, he has been a Professor with Nagoya University. Since 2021, he has been an Emeritus Professor. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and IEICE. He was awarded 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.