CROCODILE: Crop-based Contrastive Discriminative Learning for Enhancing Explainability of End-to-End Driving Models

Chenkai ZHANG, Daisuke DEGUCHI, Jialei CHEN, Zhenzhen QUAN, and Hiroshi MURASE

Nagoya University, Nagoya Aichi, Japan, zhang1354558057@gmail.com

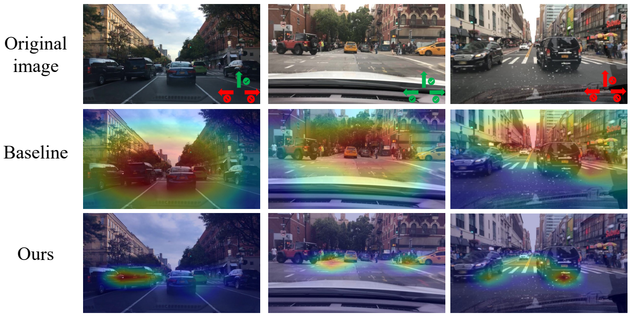
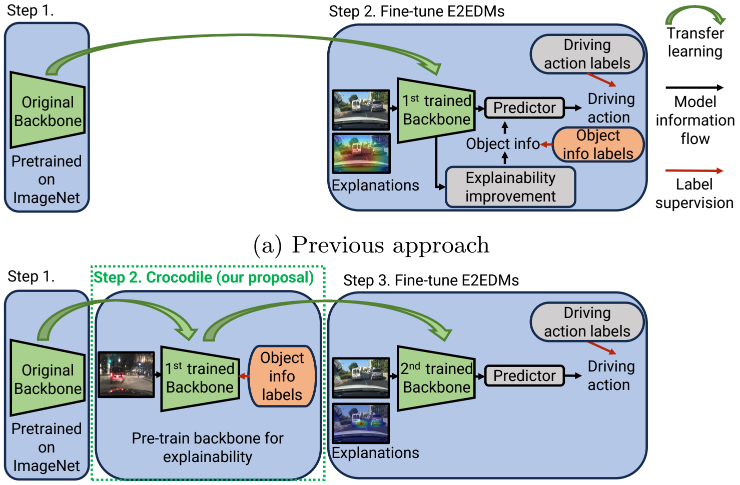


Fig. 1: In the original image, the green arrow indicates the driving action is available, the red arrow indicates that it is not. The heatmap indicates the importance of each pixel in the prediction. Our approach could help E2EDMs better focus on semantically meaningful visual features than baseline.

Abstract. In autonomous driving, visual features play a crucial role. End-to-end driving models (E2EDMs) extract numerous visual features from the driving environment to solve driving tasks. However, these visual features are often difficult for humans to understand, leading to explainability issues. This study aims to improve the explainability of E2EDMs by enhancing their ability to extract semantically meaningful and driving-related visual features, like vehicles, pedestrians, and traffic signals. The training process of E2EDMs involves leveraging a backbone that is pre-trained on large datasets and subsequently fine-tuned for driving tasks. To address the explainability issue of E2EDMs, previous studies have designed complex E2EDMs during the fine-tuning stage. In this paper, we enhance the explainability by improving the backbone's ability to recognize driving-related features, *i.e.*, object features. We propose **CRO**p-based **CO**ntrastive **DI**scriminative **LE**arning (**CROCODILE**), an additional pre-training method for the backbone. CROCODILE improves the backbone's ability to preserve driving-related features while



(a) Previous approach

(b) Our approach

Fig. 2: The comparison of the previous approach and our approach, the previous approach enhances explainability during the fine-tuning stage. On the other hand, we enhance explainability in an additional pre-training stage.

suppressing irrelevant features. Then, during fine-tuning, only driving-related features will be used for driving action prediction, thereby achieving high explainability. In addition, CROCODILE eliminates the need for complex structures in the fine-tuning stage.

1 Introduction

End-to-end driving models (E2EDMs) are the most popular autonomous driving models, they automatically extract and select visual features directly from the driving environment. These models are capable of learning optimal visual features tailored to specific driving tasks, resulting in higher prediction accuracy. However, these extracted visual features are difficult for humans to understand, leading to explainability challenges. This study aims to improve the explainability of E2EDMs by enhancing their ability to extract more semantically meaningful and driving-related visual features, making the decision-making process more transparent and understandable.

Explanation methods are used to generate explanations for E2EDMs [6, 15, 2]. There are textual-based [1, 22, 11] and visual-based explanation methods [26, 17, 21], the former generates natural language to explain why the driving models perform a specific driving action, and the latter uses visual information, i.e., images to offer intuitive explanations. Visual-based explanations are particularly advantageous in telling what and where are the responsible features for the driving action, thus in this paper, we focus on visual-based explanations. Among various visual-based explanation methods, attribution-based methods [6, 28, 2] are most prevalent, they calculate the importance score of each input element in the model's prediction. As shown in Fig. 1, since the basis of the human recognition system lies in objects, the heatmap that highlights driving-related object features is more persuasive.

Like many downstream tasks, the training of E2EDMs is also based on fine-tuning a pre-trained backbone [20, 16, 3]. The purpose of pre-training the backbone is to prepare a feature extractor capable of processing images of the driving environment. Therefore, during the fine-tuning stage, the E2EDMs could use the extracted features to predict driving actions. In other words, the features extracted by the backbone have a significant impact on the prediction method of the E2EDMs, i.e., the explainability of E2EDMs.

As shown in Fig. 2a, previous studies focused on enhancing explainability during the fine-tuning stage of E2EDMs. Specifically, they added an object detection module after the backbone, which required the E2EDMs to use object information for driving action prediction, thereby enhancing explainability. However, such a side task requires significant modifications to the architecture, deviating from their inherent end-to-end nature. Additionally, this side task demands auxiliary data, specifically, labels for object information. This requirement imposes stricter demands on the datasets in fine-tuning, as they must include not only labels for driving actions but also for object information [21, 25, 12].

To address this, we take a different path. As a visual feature extractor, the backbone plays a crucial role in the explainability of E2EDMs. Therefore, we improve explainability by enhancing the backbone's ability to process visual features through an additional pre-training stage. As shown in Fig. 2b, between the original pre-training stage and the fine-tuning stage, we introduce **CRO**p-based **CO**ntrastive **DI**scriminative **LE**arning (**CROCODILE**), CROCODILE pre-trains the backbone to preserve the driving-related features and suppress the irrelevant features. During fine-tuning, only driving-related features will be used for driving action prediction (as shown in Fig. 1), achieving high explainability.

The contributions of this paper are:

- The novelty of this paper lies in shifting the focus of enhancing the explainability of E2EDMs from the fine-tuning stage to the pre-training stage, which allows us to maintain the simplicity of the E2EDMs.
- CROCODILE decouples the simultaneous requirements for object and driving information. Specifically, we first enhance explainability on a dataset containing object information, then fine-tune driving tasks on another dataset.
- Our experiments demonstrate that CROCODILE is effective across different backbones and E2EDMs, then analyze this effectiveness in ablation study.

2 Related Work

2.1 Contrastive learning methods

Contrastive learning methods utilize positive and negative sample pairs to train the backbones [19, 14, 4]. He et al. [7] introduced a dynamic dictionary and a

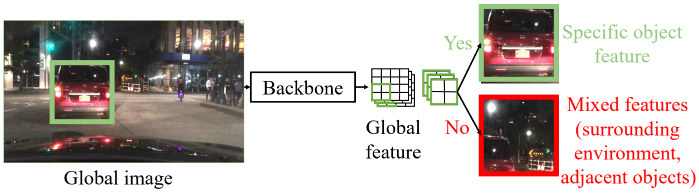


Fig. 3: Inside the global feature, we could locate the local feature corresponding to the driving-related object (green box). This important local feature should not be compromised by the unimportant background features (red box).

momentum-updated encoder to train the backbones to learn representations for image classification tasks. In object detection, contrastive learning also plays a crucial role. Zhang et al. [27] introduced Contrastive DeNoising Training, which aids the model in distinguishing between relevant objects and irrelevant background information, thereby stabilizing training and accelerating convergence.

Although contrastive learning methods have been successfully employed to enhance predictive accuracy in their respective tasks, their application in improving explainability has been limited. We believe contrastive learning methods can significantly improve explainability by guiding the discrimination between driving-related features and those that are not. Therefore, we deviate from traditional uses of contrastive learning and focus on enhancing explainability.

2.2 Approaches for enhancing the explainability of E2EDMs

Leveraging object information to enhance the explainability of E2EDMs has become a mainstream approach in the field. For instance, Wang et al. [17] utilized object features to predict driving actions. Xu et al. [21] developed a multi-task model that incorporates object labels, while Zhang et al. [25] introduced an Objectification Branch (OB) into E2EDMs to improve explainability. However, these approaches complicate the fine-tuning process and compromise the end-to-end architecture since they require E2EDMs to solve the object detection and driving tasks simultaneously. Moreover, these approaches impose strict requirements on datasets, demanding that each driving scenario include both object and driving information, which forces researchers to propose additional datasets to meet the requirements, limiting the practical applications [12, 21, 25].

To address this problem, we propose a novel method to enhance the explainability of E2EDMs. We separate the enhancement of explainability from the fine-tuning stage, eliminating the need for object detection structures.

3 Proposed Approach

3.1 Basic Idea behind our approach

If the backbone can only extract driving-related features and ignore irrelevant features, then during the fine-tuning stage of E2EDMs, only driving-related fea-

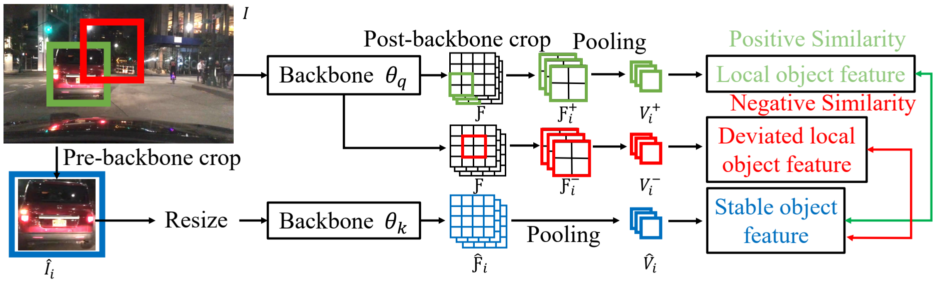


Fig. 4: The overview of **CROCODILE**. In the training process, there are two kinds of crops: pre-backbone crop and post-backbone crop. The pre-backbone crop cuts the driving-related object $\hat{I}_i$ out of the input image I, and then processes $\hat{I}_i$ with the backbone θ_k, obtaining pure driving-related object features $\hat{\mathcal{F}}_i$. The post-backbone crop cuts the global feature $\mathcal{F}$, which is obtained by processing the input image with the backbone θ_q. From $\mathcal{F}$, the post-backbone crop cuts out $\mathcal{F}_i^+$ and $\mathcal{F}_i^-$. We consider $\mathcal{F}_i^+$ and $\hat{\mathcal{F}}_i$ as positive pairs, $\mathcal{F}_i^-$ and $\hat{\mathcal{F}}_i$ are negative pairs. After training, the backbone θ_q is used for future fine-tuning.

tures will be used for driving action prediction, thereby achieving high explainability. Therefore, in this paper, we propose CROCODILE to further pre-train the backbone to endow it with this capability. Existing pre-trained backbones could process object features in images. Therefore, in this paper, we focus on further enhancing the backbone's ability to preserve driving-related object features and suppress irrelevant features. As shown in Fig. 2b, our approach further trains the pre-trained backbones [5], rather than training from scratch.

During the additional pre-training process for explainability, CROCODILE ensures that the driving-related object features are not compromised by the irrelevant features. For example, in the global image shown in Fig. 3, an object is located within a green box. After the backbone processes this global image, we obtain a global feature. Within this global feature, the local features corresponding to the object should still represent the specific object (green box), rather than being mixed with features from the surrounding environment (red box). This capability is fundamental for using the driving-related object features for driving tasks and is a prerequisite for the E2EDM's explainability.

3.2 Crop-based Contrastive Discriminative Learning

To study how the driving-related object feature is affected by the driving environment during the backbone's transformation, we perform two crops: pre-backbone crop and post-backbone crop. As shown in Fig. 4, the pre-backbone crop happens before the backbone's transformation, resulting in a pure object feature that contains no environmental information. On the other hand, the post-backbone crop happens after the backbone's transformation, resulting in an impure object feature that inevitably includes environmental information. Overall, these two types of crops result in a pure object feature and an impure object feature, we use contrastive discriminative learning to eliminate the difference between these two features by denoting them as the positive pair, thereby enhancing the backbone's ability to preserve the features of the driving-related objects from being mixed with the environmental features.

For each driving scenario, we prepare an image with the bounding box information of the driving-related object. In this paper, we describe a bounding box using the coordinates of its center point, width, and height. As shown in Fig. 4, we crop the red car from the input image as

$$\hat{I}_i = I(x_i, y_i, w_i, h_i), \tag{1}$$

where I is the input image, $\hat{I}_i$ is the i-th object image cropped (pre-backbone) from I. We use the coordinates of the object bounding box center point (x_i, y_i), along with width w_i and height h_i to crop $\hat{I}_i$ from I.

We then resize $\hat{I}_i$ to match the size of I. Then, I and $\hat{I}_i$ are processed by backbones θ_q and θ_k respectively, as

$$\mathcal{F} = \theta_q(I), \quad \hat{\mathcal{F}}_i = \theta_k(\hat{I}_i), \tag{2}$$

this process yields $\mathcal{F}$ representing the entire driving environment and $\hat{\mathcal{F}}_i$ representing the pure 2D object feature that is most related to driving tasks.

Within $\mathcal{F}$, we locate $\mathcal{F}_i^+$, the impure 2D object feature that is most related to driving tasks. We calculate the bounding box information of $\mathcal{F}_i^+$ as

$$\alpha_x = \frac{\widetilde{W}}{W}, \alpha_y = \frac{\widetilde{H}}{H} \tag{3}$$

$$\mathcal{F}_i^+ = \mathcal{F}(\alpha_x x_i, \alpha_y y_i, \alpha_x w_i, \alpha_y h_i), \tag{4}$$

where W and H are the width and height of the input image, $\widetilde{W}$ and $\widetilde{H}$ are the width and height of the feature map. $\mathcal{F}_i^+$ is cropped (post-backbone) from $\mathcal{F}$.

Similar to general contrastive discriminative learning [7, 4, 27], our approach involves positive and negative samples. We consider $\mathcal{F}_i^+$ and $\hat{\mathcal{F}}_i$ as a positive sample pair to guide the backbone to preserve the features of the driving-related object. On the other hand, to ensure that the driving-related object features are not compromised by environmental features, we design negative samples to teach the backbone to distinguish between $\hat{\mathcal{F}}_i$ and the surrounding mixed feature. We define $\mathcal{F}_i^-$, the 2D surrounding mixed feature, by keeping the size of bounding box unchanged while randomly deviating the center point of the $\mathcal{F}_i^+$ as

$$\mathcal{F}_i^- = \mathcal{F}(\alpha_x x_i \pm \epsilon * \alpha_x w_i, \alpha_y y_i \pm \epsilon * \alpha_y h_i, \alpha_x w_i, \alpha_y h_i). \ \epsilon \in [0.25, 0.5] \tag{5}$$

We apply global average pooling (GAP) to $\hat{\mathcal{F}}_i$, $\mathcal{F}_i^+$, and $\mathcal{F}_i^-$ to obtain corresponding vectors: $\hat{V}_i$, V_i^+, and V_i^-. We use these vectors to calculate the cosine similarity of the positive pair and negative pair, and then define the loss as

$$\mathcal{L} = 2 + S_i^- - S_i^+ = 2 + \frac{V_i^- \cdot \hat{V}_i}{\|V_i^-\| \|\hat{V}_i\|} - \frac{V_i^+ \cdot \hat{V}_i}{\|V_i^+\| \|\hat{V}_i\|}, \tag{6}$$

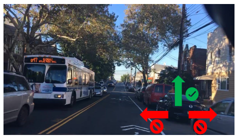


Fig. 5: A typical scene in the dataset.

where S_i^+, S_i^- are the cosine similarities for the positive and negative pairs. The range of $\mathcal{L}$ is $0 \sim 2$. As a positive pair, V_i^+ should be close to $\hat{V}_i$; as a negative pair, the V_i^- should diverge from $\hat{V}_i$. Therefore, as the learning target of positive and negative pairs, $\hat{\mathcal{F}}_i$ has to be stable to facilitate training and convergence. To achieve this, we adopt a momentum update method [7] where backbone θ_q is normally trained by back-propagation, while backbone θ_k learns from θ_q as

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{7}$$

where the momentum coefficient m is set to 0.999.

4 Experiments

4.1 Dataset

In this paper, we use two driving scene datasets. The first kind contains object information labels, which are used to train the backbones with CROCODILE. The second kind contains driving action labels, which are used to fine-tune E2EDMs.

Additionally pre-train the backbone: In the BDD-100K dataset [23], there is a collection gathered for object-tracking tasks. It comprises videos shot from a driver's perspective, each frame is annotated with the location of every object within the scene. The collection provides approximately 200K images.

Fine-tune the E2EDMs: In the BDD-3AA (3 Available Actions) [26] dataset, which annotates each scenario with the availabilities of three driving actions: acceleration, left steering, and right steering. The BDD-3AA dataset considers the driving task as a multi-label classification problem. 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 these three driving actions. For example, in a typical scene depicted in Fig. 5, the ground truth is represented as $A = [1, 1, 0]^T$, where 1 signifies an available action and 0 is an unavailable one. We utilized the macro F1 score to evaluate prediction accuracy as

$$Macro\ F_1 = \frac{F_1(\hat{A}_a, A_a) + F_1(\hat{A}_l, A_l) + F_1(\hat{A}_r, A_r)}{3}, \tag{8}$$

where A_a, A_l, A_r are the acceleration, steering left, and steering right actions, the A and $\hat{A}$ denote the ground truth label and the prediction result.

4.2 The Subjective Persuasibility Evaluation Method

As the persuasibility evaluation method for explanations proposed in [25], we gathered 5 participants who possess driver's licenses and assessed their satisfaction level with the explanations. We show the heatmaps to participants, each heatmap is evaluated by 5 participants. Participants rate each heatmap from 1 to 5, with 1 being low persuasibility and 5 being high persuasibility, and then we calculate the average value as the final score for this heatmap.

4.3 The Objective Persuasibility evaluation method

The objectification degree OD [24] represents the extent to which driving-related objects are utilized in the explanations. Given that the human recognition system relies on objects, the objectification degree determines the persuasibility of the explanation. Without using humans as participants, this method objectively evaluates the explanations generated by E2EDMs as

$$OD = \frac{\sum_{p \in O_{all}} Imp(p)}{\sum_{p} Imp(p)}, \tag{9}$$

where $Imp(p)$ represents the importance score of a pixel in the explanations, O_{all} is the mask of all objects' areas. $\sum_{p \in O_{all}} Imp(p)$ represents the summation of all pixels' importance scores inside the object area, $\sum_{p} Imp(p)$ represents the summation of all pixels' importance scores. The OD represents the proportion of the object's features among all the features important for driving actions.

4.4 Implementation details

In the driving scenario, the biggest object is typically closest to the ego vehicle and thus has a significant impact on driving decisions. Therefore, we consider the biggest object in the images as the most driving-related object, and the training of CROCODILE is focused on this biggest object. The bounding box information for all objects, including the largest object, is annotated.

In this paper, the backbones in E2EDMs are pre-trained. As shown in Fig. 2b, there are two training approaches for these backbones. In the previous approaches, the backbones are pre-trained on ImageNet [5]. On the other hand, in our approach, the backbones pre-trained on ImageNet will be further pre-trained with CROCODILE. Then, the E2EDMs utilize the trained backbones and undergo overall fine-tuning on the BDD-3AA dataset [26]. We apply 5-fold cross-validation to train each E2EDM for 50 epochs and evaluate the average accuracy on corresponding test datasets. As shown in Fig. 2b, there are also two fine-tuning approaches for E2EDMs. In the previous studies [21, 25, 12], the E2EDMs are fine-tuned with the driving action labels and the object info labels. On the other hand, in our approach, the E2EDMs are fine-tuned with only the driving action labels. For the training of backbones and E2EDMs, the Adam optimizer is utilized with a weight decay of 1×10^{-4} and a learning rate of 0.001.

Table 1: All methods and their configurations in Section 5.1. The names have a certain pattern: $a - b - c - d$, a is the backbone trained status, and it could be $Ours$ (additionally pre-trained by CROCODILE), $FRCNN$ (additionally pre-trained by Faster RCNN), or omitted (no additional pre-training); b is the backbone name; c is the E2EDM name; d is whether there is an objectification branch (OB) during fine-tuning, and it could be O (OB) or omitted (no OB).

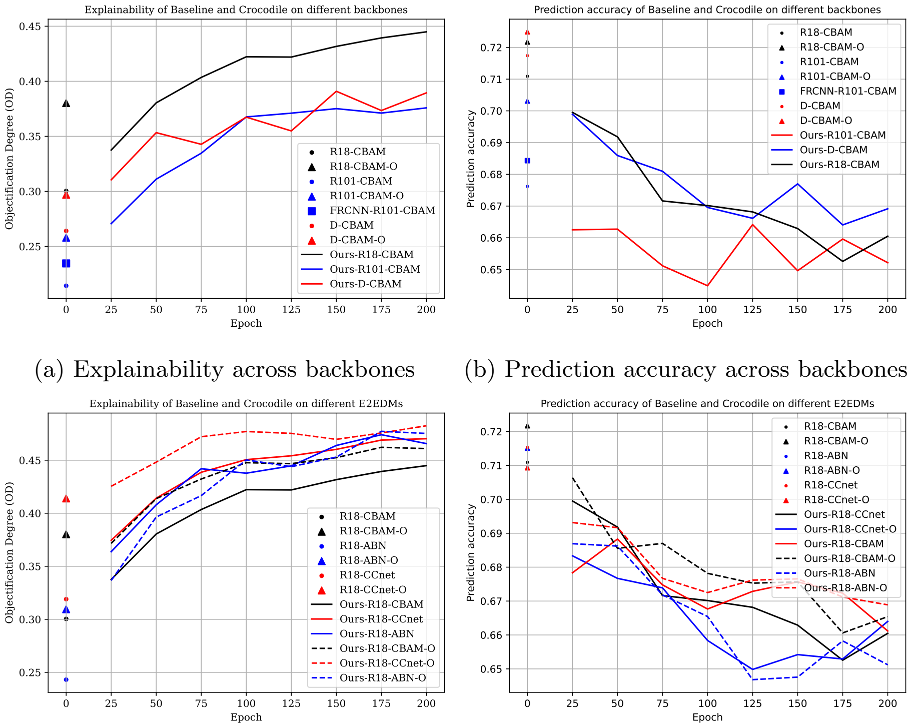
Method	Backbone	E2EDM	OB [25]	CROCODILE
R18-CBAM	ResNet18[8]	CBAM [18]		
R18-CBAM-O			✓	
Ours-R18-CBAM				✓
R101-CBAM	ResNet101[8]			
R101-CBAM-O			✓	
FRCNN-R101-CBAM				
Ours-R101-CBAM				✓
D-CBAM	DenseNet201[9]			
D-CBAM-O			✓	
Ours-D-CBAM				✓

Based on previous research [26], explanations for the E2EDMs' should based on the high-level features used to predict the driving action. Therefore, all E2EDMs in this paper are integrated with an attention mechanism [13, 18, 10] applied to these high-level features, allowing us to generate faithful explanations by overlaying the attention mask on the input images.

5 Experimental Results and Discussion

To verify the effectiveness of CROCODILE across different backbones and E2EDMs, we conducted two sets of experiments, comparing each with the corresponding baselines. In the first set of experiments, we train multiple backbones using the CROCODILE and then fine-tune the same E2EDM to present the effectiveness of CROCODILE on various backbones. Based on the first set of experiments, we identify the backbone on which CROCODILE performs best. In the second set of experiments, we fine-tune various E2EDMs on this backbone to present the effectiveness of CROCODILE on various E2EDMs. Furthermore, we present a detailed ablation study to analyze the source of the effectiveness.

Most results are averaged over five runs to minimize the randomness of our experimental results. For instance, each backbone trained with CROCODILE is trained five times, and an E2EDM is fine-tuned on each backbone. The average accuracy and explainability of the five E2EDMs are then presented as the final result. However, due to the high cost of some experiments, we fine-tuned the E2EDM on only one backbone, and we will explicitly mention such cases.



(a) Explainability across backbones (b) Prediction accuracy across backbones

(c) Explainability across E2EDMs (d) Prediction accuracy across E2EDMs

Fig. 6: CROCODILE and baseline across backbones and E2EDMs.

5.1 The effectiveness of CROCODILE across different backbones

For the baseline, ZHANG et al. [25] proposed the Objectification Branch (OB), which uses object information during fine-tuning to enhance the explainability of E2EDMs. Since their E2EDMs used the Convolutional Block Attention Mechanism [18], we refer to their two E2EDMs as CBAM and CBAM-O, the CBAM-O has OB and CBAM does not.

We train a backbone with CROCODILE for 200 epochs, saving the backbone every 25 epochs, resulting in eight backbones with different levels of training. Then, we fine-tune the CBAM based on this series, resulting in 8 E2EDMs. To demonstrate the effectiveness of CROCODILE across different backbones, as shown in Table 1, we train 3 series of E2EDMs and their corresponding baselines. In addition, we add another baseline for Ours-R101-CBAM. Since CROCODILE enhances the explainability of the E2EDM by enabling the backbone to preserve important object features, we further prove the effectiveness of CROCODILE by comparing it with a traditional object detection task. In Ours-R101-CBAM, we replace CROCODILE with an object detection task. Specifically, after obtaining the backbone pre-trained on ImageNet, the backbone is trained on the BDD-

Table 2: The subjective persuasibility of explanations.

Method	R101-CBAM	R101-CBAM-O	Ours-R101-CBAM
Subjective persuasibility	2.48	2.70	**3.44**

Table 3: All methods and their configurations in Section 5.2. Compared to Table 1, the backbone is ResNet18, and we present more E2EDMs by replacing CBAM [18] with ABN [13] and CCnet [10]. In addition, to investigate the impact on explainability when combining CROCODILE with the baseline [25], we train 3 series of E2EDMs, *e.g.*, for Ours-R18-CBAM, there is Ours-R18-CBAM-O.

Method	Backbone	E2EDM	OB [25]	CROCODILE
R18-CBAM	ResNet18 [8]	CBAM [18]		
R18-CBAM-O			✓	
Ours-R18-CBAM				✓
Ours-R18-CBAM-O			✓	✓
R18-CCnet		CCnet [10]		
R18-CCnet-O			✓	
Ours-R18-CCnet				✓
Ours-R18-CCnet-O			✓	✓
R18-ABN		ABN [13]		
R18-ABN-O			✓	
Ours-R18-ABN				✓
Ours-R18-ABN-O			✓	✓

100K dataset for object detection using Faster RCNN. Then, we fine-tune the CBAM based on this backbone and refer to it as FRCNN-R101-CBAM.

As shown in Fig. 6a, we present the objective evaluation of the explainability of the aforementioned E2EDMs. The horizontal axis represents training duration, and the vertical axis represents the OD of each E2EDM, which indicates how much object feature information the E2EDM uses to make driving action predictions. Since the human cognitive system is object-based, a higher OD indicates higher explainability. Based on the performance of CROCODILE on 3 backbones, we could see that for any backbone, as the training duration increases, the explainability of the E2EDM gradually improves and outperforms all baselines. To further prove the effectiveness of CROCODILE, we evaluated the human subjective persuasibility of the explanations from Ours-R101-CBAM (200-th epoch), R101-CBAM, and R101-CBAM-O. Due to the high cost of this experiment, these E2EDMs are trained only once. As shown in Table 2, the explanations generated by Ours-R101-CBAM are easier to understand.

In addition, we can see that Ours-R101-CBAM consistently outperforms those of FRCNN-R101-CBAM. This indicates that the object detection task cannot replace CROCODILE in enhancing the explainability of the E2EDMs.

Table 4: Summary of all methods and their configurations in Section 5.3, there are three components in CROCODILE, DRF, CP, and Negative. For DRF, the BO represents the Biggest Object, RO represents the Random Object, and RC represents the Random Crop. For CP, the Crop represents using the original pre-backbone and post-backbone crops to make positive and negative pairs, the CDN represents using the CDN in [27] to make positive and negative pairs. For Negative, the check mark and the cross mark represent whether we consider the similarity of negative pairs in Eq. 6.

Methods	Backbone	E2EDM	CROCODILE components					
			DRF			CP		Negative
			BO	RO	RC	Crop	CDN	
Ours-R18-CCnet	ResNet18 [8]	CCnet [10]	✓			✓		✓
RO-R18-CCnet				✓		✓		✓
RC-R18-CCnet					✓	✓		✓
CDN-R18-CCnet			✓				✓	✓
w/o Negative-R18-CCnet			✓			✓		

This outcome may be expected because what benefits the E2EDMs is not a backbone that excels at extracting all object information but the important objects.

Finally, we analyze the prediction accuracy of all E2EDMs. As shown in Fig. 6b, prediction accuracy decreases as the training duration increases. There is a typical trade-off between prediction accuracy and explainability.

5.2 The effectiveness of CROCODILE across different E2EDMs

In Section 5.1, we found that CROCODILE performed best on ResNet18, thus in Table 3, we fine-tune E2EDMs and their corresponding baselines on ResNet18.

As shown in Fig. 6c, for each E2EDM, the performance of CROCODILE consistently surpasses their corresponding baselines. Furthermore, when combining CROCODILE with the baseline methods, the Ours-R18-CBAM-O and Ours-R18-CCnet-O achieve superior performance than Ours-R18-CBAM and Ours-R18-CCnet-O, the Ours-R18-ABN-O achieve similar performance with Ours-R18-ABN. These results confirm the effectiveness of the CROCODILE. Next, we compare the heatmaps generated by different E2EDMs to intuitively understand the differences between CROCODILE and the baselines. As shown in Fig. 7, the E2EDMs from the CROCODILE have a stronger ability to utilize important object features compared to the baselines.

Finally, in Fig. 6d, as same as it does in Fig. 6b, prediction accuracy decreases as the training duration increases.

5.3 Ablation Study

We remove or replace certain components of the CROCODILE and observe whether the modified CROCODILE remains effective. Since we found that our

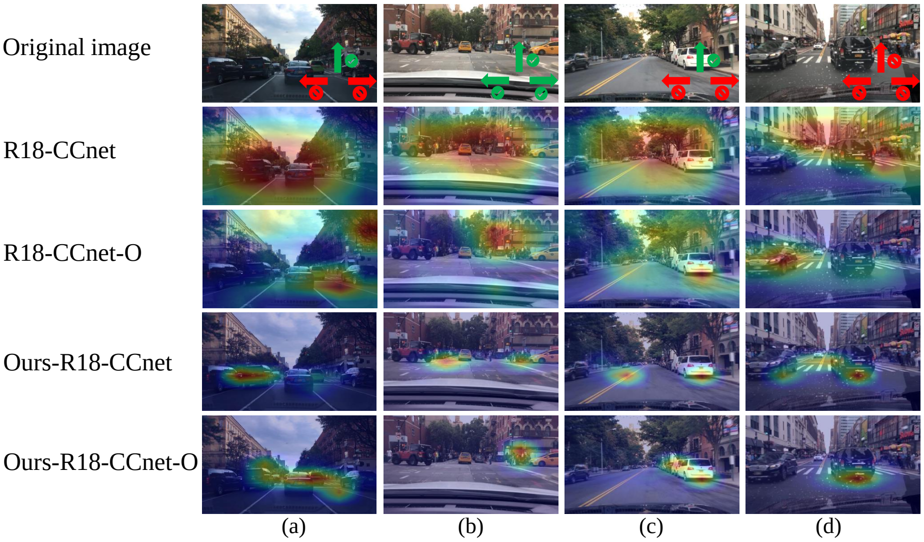


Fig. 7: Columns represent different driving scenes, rows represent E2EDMs.

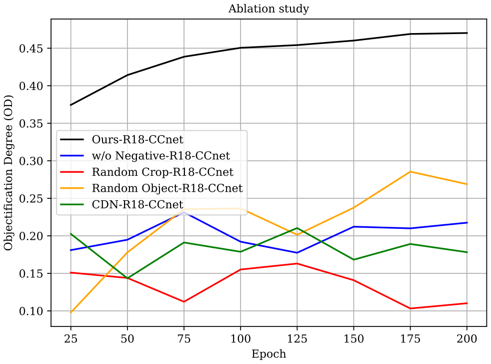


Fig. 8: The explainability of CROCODILE and its variations.

approach performs best with ResNet18 and CCnet, we first train ResNet18 using different versions of CROCODILE and then fine-tune it on CCnet. As shown in Table 4, we analyze the effectiveness of various versions of the CROCODILE.

- Driving-Related Feature (**DRF**): As introduced in Section 4.4, CROCODILE chooses the biggest object (**BO**) as the driving-related feature. We make two different modifications. **1.** We randomly select an object from all the objects. **2.** Instead of selecting objects, we randomly crop a region from the image.
- Contrastive Pairs (**CP**): CROCODILE determines the positive and negative samples in contrastive learning through two types of crops. To observe the impact of the crops, we replace it with another method to determine the

positive and negative samples. In previous research, ZHANG et al. [27] propose Contrastive DeNoising (CDN). CDN randomly alters the coordinates of the top-left and bottom-right corners of an object's bounding box. A new bounding box with small changes is used as a positive sample, while a new bounding box with large changes is used as a negative sample, we replace the crops with CDN for selecting positive and negative samples.
- **Negative**: As introduced in Eq. 6, CROCODILE not only considers the similarity of the positive pairs but also considers negative pairs. We make a modified CROCODILE by only considering the similarity of positive pairs.

As shown in Fig. 8, for the first component, changing the BO to RO or RC significantly weakens the explainability of the E2EDM. More specifically, the RC results in worse explainability than the RO. This implies that object information, and particularly important object information, is crucial for explainability. For the second component, using CDN to select positive and negative samples significantly weakens the explainability. This indicates that the crops in CROCODILE is crucial. For the third component, without negative samples, the explainability is significantly weakened. This aligns well with the understanding of contrastive learning: having only positive samples without negative samples leads to the backbone taking shortcuts, *i.e.*, output the same feature for any image. All explanations discussed in Section 5 are shown in supplementary materials.

6 Conclusion

In this paper, we proposed CROCODILE, a method that enhances the explainability of E2EDMs by adding an additional pre-training stage for the backbone. CROCODILE determines the positive and negative samples for contrastive learning through two types of crops, enabling the backbone to better process the driving environment image. Specifically, the important object features in the image are well-preserved, without being mixed with background information. This provides a foundation for fine-tuning the E2EDM, allowing it to only use important object features for driving action prediction, thereby improving explainability.

Although our method successfully enhances the explainability of E2EDMs, we observed a decline in prediction accuracy, showing the trade-off between these two aspects. Addressing the challenge of simultaneously improving both prediction accuracy and explainability is a key focus of our future research.

Acknowledgement

This work was supported by JST SPRING JPMJSP2125, JSPS KAKENHI Grant Number 23K28164, and JST CREST Grant Number JPMJCR22D1. The author Chenkai Zhang would like to take this opportunity to thank the "Interdisciplinary Frontier Next-Generation Researcher Program of the Tokai Higher Education and Research System."

References

1. Ben-Younes, H., Zablocki, É., Pérez, P., Cord, M.: Driving behavior explanation with multi-level fusion. Pattern Recognition **123**, 108421 (2022)
2. Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., Zieba, K.: Visualbackprop: visualizing cnns for autonomous driving. arXiv preprint arXiv:1611.05418 **2**, 1–2 (2016)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
10. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 603–612 (2019)
11. Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., Liu, J.: Adapt: Action-aware driving caption transformer. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7554–7561. IEEE (2023)
12. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–578 (2018)
13. Mori, K., Fukui, H., Murase, T., Hirakawa, T., Yamashita, T., Fujiyoshi, H.: Visual explanation by attention branch network for end-to-end learning-based self-driving. In: 2019 IEEE intelligent vehicles symposium (IV). pp. 1577–1582. IEEE (2019)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. Ras, G., Xie, N., Van Gerven, M., Doran, D.: Explainable deep learning: A field guide for the uninitiated. Journal of Artificial Intelligence Research **73**, 329–396 (2022)
16. Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Muhammad, N.: A survey of end-to-end driving: Architectures and training methods. IEEE Transactions on Neural Networks and Learning Systems **33**(4), 1364–1384 (2020)
17. Wang, D., Devin, C., Cai, Q.Z., Yu, F., Darrell, T.: Deep object-centric policies for autonomous driving. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 8853–8859. IEEE (2019)

18. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
19. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
20. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2174–2182 (2017)
21. Xu, Y., Yang, X., Gong, L., Lin, H.C., Wu, T.Y., Li, Y., Vasconcelos, N.: Explainable object-induced action decision for autonomous vehicles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9523–9532 (2020)
22. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)
23. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
24. Zhang, C., Deguchi, D., Chen, J., Murase, H.: Toward explainable end-to-end driving models via simplified objectification constraints. IEEE Transactions on Intelligent Transportation Systems (2024)
25. Zhang, C., Deguchi, D., Murase, H.: Refined objectification for improving end-to-end driving model explanation persuasibility. In: 2023 IEEE Intelligent Vehicles Symposium (IV). pp. 1–6. IEEE (2023)
26. Zhang, C., Deguchi, D., Okafuji, Y., Murase, H.: More persuasive explanation method for end-to-end driving models. IEEE Access **11**, 4270–4282 (2023)
27. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
28. Zhang, Y., Tiňo, P., Leonardis, A., Tang, K.: A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence **5**(5), 726–742 (2021)