Enhancing Explainability of End-to-End Autonomous Driving Models through additional fine-tuning

Chenkai ZHANG[1,a)] Daisuke Deguchi[1,b)] Jialei Chen[1,c)] Hiroshi Murase[1,d)]

Abstract

In autonomous driving, end-to-end driving models (E2EDMs) have become increasingly popular due to their exceptional predictive performance. This performance stems from leveraging a backbone that is pre-trained on large datasets and then fine-tuned for driving tasks. However, the opaque nature of these E2EDMs presents a challenge to explainability, thus visual explanations are created to shed light on the decision-making process of E2EDMs. Currently, these E2EDMs have been made more explainable by fine-tuning the driving tasks along with a side task designed for explainability. This side task for explainability requires complex structures with auxiliary data, such as the location of objects.

In this paper, we argue that a more effective approach to enhancing the explainability of E2EDMs is to design a fine-tuning specifically focused on explainability. We propose **CRO**p-based **CO**ntrastive **DI**scriminative **LE**arning (**CROCODILE**) to enhance a backbone's capability to accurately identify object features during this additional fine-tuning for explainability. By adopting **CROCODILE**, we can develop more explainable E2EDMs without relying on auxiliary data or complex structures in the fine-tuning for driving tasks. Our experimental results confirm that our approach enhances the explainability of E2EDMs.

1. Introduction

Autonomous driving models [1] are primarily categorized into two types: modular pipeline systems [2] and end-to-end driving models (E2EDMs) [3]. Modular pipeline systems are known for their higher explainability but face limitations in predictive accuracy. In contrast, E2EDMs offer higher accuracy but suffer from low explainability due to their "black box" nature. This accuracy-explainability trade-off presents a significant challenge in the field of autonomous driving [4].

As deep learning technology rapidly evolves, researchers are focusing more on enhancing the explainability of E2EDMs [3], [5], [6] to develop autonomous driving models that are both accurate and easily understandable. In explaining E2EDMs, explanation methods are used to provide explanations for upcoming observations [7], [8], [9].

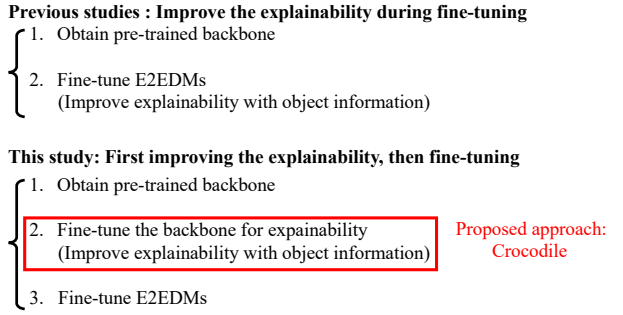

Fig. 1 The comparison of our proposed method and previous method.

Numerous strategies [11], [12], [13] enhance the explainability of E2EDMs by incorporating side tasks that need auxiliary data and complex structures during the fine-tuning for driving tasks. However, this approach can necessitate substantial modifications to the E2EDMs' architecture, deviating from their inherent end-to-end nature. Such modifications can complicate the training process and affect the core design of E2EDMs.

In response to these challenges, we aim to improve the explainability of purely E2EDMs without compromising their fundamental design. As depicted in Fig. 1, instead of improving explainability in the fine-tuning for driving tasks, we add an additional fine-tuning process specifically designed for improving explainability, it is situated between pre-training and fine-tuning, it concentrates on training the E2EDMs' backbone to better recognize objects. Since the human recognition system is based on objects, the capability to use objects to predict driving actions determines the explainability of E2EDMs. The novelty of this paper is separating the explainability-enhancing training from the driving-task training, allowing us to preserve the end-to-end integrity of the E2EDMs. Our experiments have shown that our approach significantly enhances the explainability of E2EDMs beyond what previous attempts have achieved.

[1] Nagoya University
[a)] zhang.chenkai.d4@s.mail.nagoya-u.ac.jp
[b)] ddeguchi@nagoya-u.jp
[c)] jialeichen2021@gmail.com
[d)] murase@nagoya-u.jp

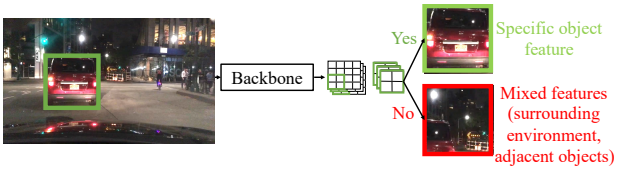

Fig. 2 The basic idea of our approach.

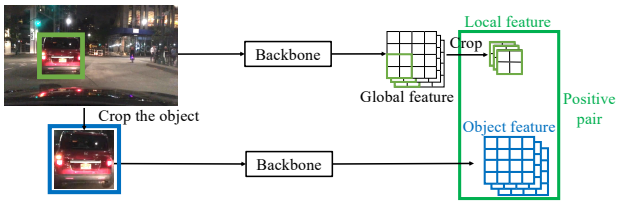

Fig. 3 The definition of positive pairs.

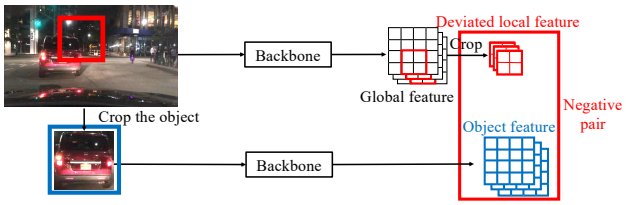

Fig. 4 The definition of negative pairs.

2. Related work

2.1 Contrastive learning approaches

Contrastive learning [14], [15], [16], [17], pivotal in self-supervised learning for image-based computer vision, emphasizes learning representations by using positive pairs and negative ones. A seminal work in this domain is "Momentum Contrast (MoCo)," [17] which innovated a dynamic dictionary, queue, and momentum updated encoder for effective contrastive learning, demonstrating significant efficacy in developing high-quality representations.

Although methods like MoCo have excelled in boosting predictive accuracy in downstream tasks, their focus has not been on enhancing model explainability. Our framework diverges from traditional contrastive learning by aiming to improve the explainability of models within learning tasks, marking a novel direction in this field.

2.2 Enhance the explainability of E2EDMs

Efforts to enhance the explainability of E2EDMs have seen varied success. Wang et al. [11] initially used object features for driving actions but observed a decrease in prediction accuracy, which they later attempted to rectify by incorporating 3D object information. Xu et al. [12] developed a multi-task model with object labels. In our prior work [13], we introduced the ROB structure to be integrated into E2EDMs to improve explainability. However, these approaches often steered away from the integral end-to-end architecture, compromising the inherent benefits of E2EDMs.

In this paper, we proposed a novel method to enhance the explainability of E2EDMs. Our method distinctly separates the enhancement of explainability from the fine-tuning phase. We focus on training the E2EDMs' backbone with object information in a phase before fine-tuning, thus enhancing explainability without compromising the core end-to-end model structure.

3. Method

3.1 The basic idea of our approach

Based on previous studies [18], [19], [20], a high-explainability E2EDM should possess a robust capability to process object information. Therefore, our goal is to enhance the backbone with this critical ability. We introduce **CRO**p-based **CO**ntrastive **DI**scriminative **LE**arning (**CROCODILE**) during an additional fine-tuning process to train the backbone. The basic idea is that the backbone must exhibit strong proficiency in processing object features within images, specifically preserving object features in their corresponding spatial locations. An example is shown in Fig. 2, where an object is located within a green box in the global image. After processing this global image through the backbone, we obtain a **global feature**. Within this global feature, the **local features** corresponding to the object's location should still represent the specific object within the green box, rather than being a mix of features from the surrounding environment or adjacent objects. We use the contrastive learning method to realize this idea, thus introducing the design of positive and negative pairs.

3.2 CROCODILE: Crop-based contrastive discriminative learning

3.2.1 The design for positive pairs

Positive pairs serve an important role in the contrastive learning method, they indicate that two features should be similar. In this paper, we denote a **local feature** and an **object feature** as positive pairs. As shown in Fig. 3, the **local feature** is cropped from the global feature, and the crop position is the position of a specific object in the input image. To obtain the **object feature**, we first crop the specific object from the input image and then input it to the backbone. Although both features target an object, the local feature inevitably includes some contextual information from the entire image, the object feature represents pure object-specific features. By using these two features as positive pairs, we train the backbone to effectively capture and retain object features even when processing the entirety of an image. This ensures that the backbone develops a robust capability to process object information.

3.2.2 The design for negative pairs

On the other hand, negative pairs indicate that two features should be dissimilar. In this paper, we denote a **deviated local feature** and the **object feature** as negative pairs. As shown in Fig. 4, the **deviated local feature** is

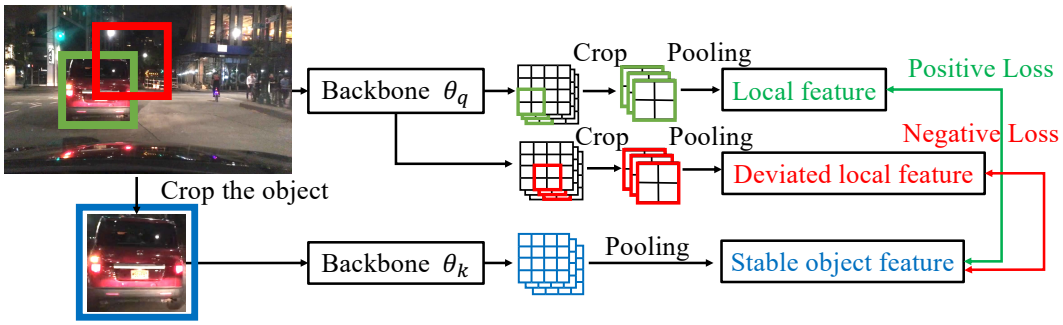

Fig. 5 The comparison of our proposed method and previous method, the backbone θ_q is the resulting backbone for future usage.

also cropped from the global feature, but the crop position is a random position near the specific object in the input image. By using these two features as negative pairs, we train the backbone to discriminate the surrounding environment or adjacent objects from the target object feature.

3.2.3 The combination of positive pairs and negative pairs

We simply combine the positive pairs and the negative pairs to obtain the final **CROCODILE**. To enhance the backbone's capability to accurately detect objects within images, we utilize the known location information of various objects (like vehicles, pedestrians, and traffic lights) in the image to define positive and negative pairs. Note, as shown in Fig. 5, the object feature serves as a teacher for the local feature and the deviated local feature. Therefore, to obtain a relatively stable object feature to train the network, the backbone θ_k is momentum updated [17] by θ_q.

Similar to previous studies that enhance explainability by fine-tuning the driving tasks along with a side task designed for explainability, our method also requires object location information for training. However, the key difference lies in our approach is where to utilize this object information. Previous research used backbones pre-trained on ImageNet [22] and then used object information during the fine-tuning of the E2EDMs. In contrast, after we obtain the pre-trained backbones, we advance the stage of using object information and train the backbone with **CROCODILE** in an additional fine-tuning specifically designed for explainability.

4. Experiment

4.1 Dataset

In this paper, we use two datasets derived from BDD-100K, each serving a specific purpose. The first dataset is used to enhance the backbone's capability in detecting and recognizing objects. The second dataset is designed to fine-tune E2EDMs [23].

4.1.1 Fine-tune the backbone with CROCODILE for explainability: The BDD-100K Dataset

In the BDD-100K dataset, this collection is specifically gathered for object-tracking tasks. As illustrated in Fig. 6,

Fig. 6 In a typical scene in the dataset, the green arrow with a check mark indicates availability, while the red arrow with a forbidden character indicates that it is not.

it comprises videos shot from a driver's perspective, capturing driving environments. Each frame in these videos is annotated with the location of every object within the scene, thus we can further pre-train the backbone. The collection provides approximately 200K images.

4.1.2 Fine-tune the E2EDMs for driving tasks: The BDD-3AA Dataset

Previous driving datasets [3], [5] primarily labeled a single driver-chosen action as correct for each scenario, overlooking the reality that drivers often have multiple valid options. This approach risked training E2EDMs with a narrow understanding of driving situations, limiting the effectiveness of their explanations.

To overcome this, we adopted the BDD-3AA (3 Available Actions) [23] dataset, which annotates each scenario with three possible actions: acceleration, left steering, and right steering. This approach transformed the driving task into a multi-label classification problem, better suited for evaluating the persuasibility of E2EDM explanations. Classification tasks like these are particularly effective for assessing the quality of explanations provided by E2EDMs.

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, left steering, and right steering. For example, in a typical scene depicted in Fig. 6, the ground truth is represented

Fig. 7 The comparison of the explanations generated by the E2EDMs trained by CROCODILE and baseline approach, respectively.

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, which involved computing the average F1 score of the three actions (acceleration, steering left, and steering right).

$$Macro\, F_1 = \frac{F_1(\hat{A}_a, A_a) + F_1(\hat{A}_l, A_l) + F_1(\hat{A}_r, A_r)}{3}, \quad (1)$$

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

4.2 Experiment method to compare our method with the previous method

To demonstrate the enhanced explainability of E2EDMs developed with our proposed method compared to those using traditional methods, we trained two E2EDMs with two approaches, the baseline approach and the CROCODILE approach. For simplification, these two E2EDMs are referred to as baseline and CROCODILE, respectively.

I. **Baseline**: This E2EDM follows the conventional training framework: 1. Obtain a pre-trained backbone. 2. Fine-tune the E2EDM with additional object information to improve explainability [13].

II. **CROCODILE**: This E2EDM follows the framework proposed in our paper: 1. Obtain a pre-trained backbone. 2. Train the backbone with additional object information to improve explainability. 3. Fine-tune the E2EDM without additional object information [13].

Both E2EDMs were subjected to the same explanation method [24] to generate explanations for observation. The subsequent sections will present these explanations, showing the enhanced explainability achieved by our approach.

5. Experiment results and discussion

An experimental method to evaluate the persuasibility of explanations is proposed in [23]. 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. We assess the participants' satisfaction level with the explanations. Participants rate the heatmap (as shown in Fig. 7) from 1 to 5, with 1 being low persuasibility and 5 being high persuasibility.

Table 1 The experimental evaluation results of explanations.

Method	Baseline	CROCODILE
Heatmap Satisfaction	2.61	**3.37**

As shown in Table. 1, the explanations generated by **CROCODILE** are more persuasive than the baseline. Specifically, Fig. 7 displays the explanations generated by **CROCODILE** and baseline. Our method distinctly demonstrates a more concentrated focus on object information within the images. This evidences that integrating object information before the fine-tuning phase of the E2EDMs substantially enhances the explainability.

6. Conclusion

This paper introduces a novel approach to improve the explainability of E2EDMs. Unlike traditional methods that depend on complex structures and additional data during the fine-tuning phase for better explainability, our method underscores the value of further pre-training (before the final fine-tuning) of the E2EDMs' backbone. This advanced pre-training results in a more proficient backbone, and when these E2EDMs are fine-tuned subsequently, they exhibit notably improved explainability.

Acknowledgment

This work was supported by JST SPRING JPMJSP2125, JSPS KAKENHI Grant Number 23H03474, 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] Yurtsever E, Lambert J, Carballo A, et al. A survey of autonomous driving: Common practices and emerging technologies[J]. IEEE access, 2020, 8: 58443-58469.

[2] Levinson J, Askeland J, Becker J, et al. Towards fully autonomous driving: Systems and algorithms[C]//2011 IEEE intelligent vehicles symposium (IV). IEEE, 2011: 163-168.

[3] Bojarski M, Del Testa D, Dworakowski D, et al. End to end learning for self-driving cars[J]. arXiv preprint arXiv:1604.07316, 2016.

[4] Lee J, Moray N. Trust, control strategies and allocation of function in human-machine systems[J]. Ergonomics, 1992, 35(10): 1243-1270.

[5] Xu H, Gao Y, Yu F, et al. End-to-end learning of driving models from large-scale video datasets[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2174-2182.

[6] Tampuu A, Matiisen T, Semikin M, et al. A survey of end-to-end driving: Architectures and training methods[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 33(4): 1364-1384.

[7] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models[J]. ACM computing surveys (CSUR), 2018, 51(5): 1-42.

[8] Ras G, Xie N, Van Gerven M, et al. Explainable deep learning: A field guide for the uninitiated[J]. Journal of Artificial Intelligence Research, 2022, 73: 329-397.

[9] Bojarski M, Choromanska A, Choromanski K, et al. Visualbackprop: visualizing cnns for autonomous driving[J]. arXiv preprint arXiv:1611.05418, 2016, 2.

[10] Gilpin L H, Bau D, Yuan B Z, et al. Explaining explanations: An overview of interpretability of machine learning[C]//2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018: 80-89.

[11] Wang D, Devin C, Cai Q Z, et al. Deep object-centric policies for autonomous driving[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 8853-8859.

[12] Xu Y, Yang X, Gong L, et al. Explainable object-induced action decision for autonomous vehicles[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9523-9532.

[13] Zhang, Chenkai, Daisuke Deguchi, and Hiroshi Murase. "Refined Objectification for Improving End-to-End Driving Model Explanation Persuasibility." 2023 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2023.

[14] Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3733-3742.

[15] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[16] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PMLR, 2020: 1597-1607.

[17] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.

[18] Chen C, Seff A, Kornhauser A, et al. Deepdriving: Learning affordance for direct perception in autonomous driving[C]//Proceedings of the IEEE international conference on computer vision. 2015: 2722-2730.

[19] Sauer A, Savinov N, Geiger A. Conditional affordance learning for driving in urban environments[C]//Conference on robot learning. PMLR, 2018: 237-252.

[20] Scholl B J. Objects and attention: The state of the art[J]. Cognition, 2001, 80(1-2): 1-46.

[21] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[22] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[23] Zhang C, Deguchi D, Okafuji Y, et al. More Persuasive Explanation Method For End-to-End Driving Models[J]. IEEE Access, 2023.

[24] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.