Traffic Scene-Informed Attribution of Autonomous Driving Decisions

Rui Shi, Tianxing Li, Yasushi Yamaguchi, Liguo Zhang

***Abstract*—Deep neural networks (DNNs) have advanced autonomous driving, but their lack of transparency remains a major obstacle to real-world application. Attribution methods, which aim to explain DNN decisions, offer a potential solution. However, existing methods, primarily designed for image classification models, often suffer from performance degradation and require specialized algorithmic adjustments when applied to the diverse models in autonomous driving. To address this challenge, we introduce a universally applicable representation of traffic scenes, forming the basis for our unified attribution method. Specifically, we leverage the first-order Taylor expansion at a specific hidden layer, *i.e.*, the product of gradients and feature maps, to represent abstract traffic scene information. This representation guides both the optimization of attribution path generation and the attribution computation, enabling consistent and effective attributions for both lane-change prediction and vision-based control models. Experiments on two distinct autonomous driving models demonstrate that our approach outperforms state-of-the-art methods in explanation accuracy and robustness, advancing the interpretability of DNN-based autonomous driving models.**

***Index Terms*—Autonomous driving, neural networks, attribution methods, explainable artificial intelligence.**

I. Introduction

THE advancement of artificial intelligence, particularly in autonomous driving technologies, presents significant potential for enhancing vehicular safety and reducing the reliance on human labor [1], [2]. Recent integration of deep neural networks (DNNs) has driven substantial progress in critical applications such as traffic object detection [3], [4], scenario understanding [5], [6], trajectory prediction [7], [8], path planning [9], [10], and end-to-end decision-making [11], [12]. However, despite numerous advancements, the inherent opacity of DNN decision-making processes remains difficult to interpret, continuing to hinder transparency in autonomous driving.

Explaining DNN models in autonomous driving requires clarifying the rationale behind their output given specific inputs. Although numerous attribution studies, predominantly focused on image classification [13]–[16], have emerged to address this need, the diversity of tasks and models in autonomous driving presents unique challenges. Autonomous driving models range from end-to-end vision systems to lane-change models relying on communication or perceptual data [17]–[19]. Directly applying image classification attribution methods to these models often yields misleading results. For instance, the Aumann-Shapley (AS) method [20], despite its known efficacy in other domains, can incorrectly attribute importance to irrelevant features in a discretionary lane-change model [19]. Although designing model-specific attribution methods could improve accuracy, the resulting variance in computation methods complicates comparative analysis. Therefore, a key challenge in understanding autonomous driving decision-making lies in developing a unified and accurate attribution computation model applicable across diverse traffic scenarios and tasks.

Unlike conventional classification tasks with relatively straightforward scenarios, traffic scenes are inherently complex and highly sensitive to subtle changes that can significantly influence decision-making [21], [22]. This complexity explains why existing attribution methods, often designed without explicit consideration of traffic scene information, struggle to maintain quality and generalize across different models in autonomous driving. Importantly, traffic scene information mentioned here is not merely data acquired at the perceptual level; rather, it represents an abstract, implicit representation of scene information processed by DNNs. By incorporating this

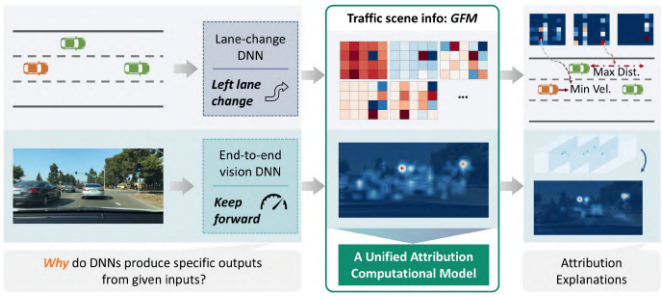


Fig. 1. Illustration of our attribution explanation. Our method extracts traffic scene information and provides a unified attribution computation model to generate explanations across different models. In the lane-change model example, the attribution explanation highlights the contributions of the ego vehicle's minimum speed and the maximum headway distance of the left-preceding vehicle (relative to its preceding vehicle) to the model's left lane-change prediction. For the vision model example, the explanation reveals that the traffic light and the white vehicle ahead are the primary factors influencing the model's decision.

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62403017, 62402021, U2233211), Beijing Natural Science Foundation (Grant No. 4244088, L243026), and Japan Society for the Promotion of Science (JSPS KAKENHI Grant No. 20H04203). (Corresponding author: Tianxing Li)

Rui Shi and Liguo Zhang are with the School of Information Science and Technology, Beijing University of Technology, Beijing 100124, China (e-mail: ruishi@bjut.edu.cn; zhangliguo@bjut.edu.cn)

Tianxing Li is with the College of Computer Science, Beijing University of Technology, Beijing 100124, China (e-mail: litianxing@bjut.edu.cn)

Yasushi Yamaguchi is with the Department of General Systems Studies, the University of Tokyo, Tokyo 153-8902, Japan (e-mail: yama@graco.c.u-tokyo.ac.jp)

Manuscript received April 19, 2021; revised August 16, 2021.

abstract information into the AS attribution computation, our method substantially improves attribution quality and enables robust application across diverse autonomous driving models, including both lane-change and vision-based models.

To effectively represent traffic scene information, our method draws inspiration from studies on Taylor expansions of DNNs [23], [24]. These studies demonstrate how DNNs, or specific hidden layers within them, can be decomposed to extract crucial information from given inputs. We utilize the first-order term of the Taylor expansion, specifically the product of gradients and feature maps (GFM) at a specific hidden layer, to provide a deep and abstract representation of the input traffic scenario. This GFM representation guides our attribution computation process.

More specifically, we employ GFM to define the integration path in the AS attribution computation. This path aligns the integral process of attribution computation with the gradual reconstruction of traffic scene information, effectively minimizing incorrect attributions. Although the endpoint of this integration path is readily available, *i.e.*, the input sample being explained, the starting point requires careful consideration. To obtain the starting point, we develop an optimization method based on the distribution of traffic scene features. This method generates a starting point that accurately represents the absence of the current scene information, thereby triggering decision changes crucial for extracting the relative importance of features. The defined integration path enables us to generate accurate attribution explanations across different model architectures in a unified form, as shown in Fig. 1. The lane-change model used is based on the work presented in [19], while the vision model is a DenseNet-based architecture as described in [25].

In summary, the primary contributions of this work are as follows:

- By abstracting traffic scene information as GFM and integrating it into the AS attribution computation, we develop a unified and highly generalizable attribution method. We demonstrate our method by applying it across diverse autonomous driving models with substantial functional differences, consistently achieving superior performance.
- Technically, we use GFM to design a context-aware attribution integration path, and determine its starting point based on the distribution of traffic scene features. This ensures that the attribution computation is informed by traffic scene throughout the entire path, enabling more accurate interpretation of model decision-making.

II. Related Work

A. Attribution Methods for Autonomous Driving

Attribution methods constitute a fundamental approach for explaining the decision-making process of DNNs. These methods assign attribution scores to input features, quantifying their relative contribution to a specific decision [26]. Attribution methods can be broadly categorized into perturbation-based, propagation-based, and other approaches. Propagation-based methods have emerged as the dominant approach due to advancements in graphics processing units (GPUs) and deep learning libraries, which have significantly improved the efficiency of forward and backward propagation within DNNs, enabling efficient computation of attributions. Our proposal, grounded in the Aumann-Shapley (AS) value, falls within the category of propagation-based methods.

Perturbation-based methods calculate attributions by introducing perturbations to input features and observing the resulting changes in the output. For instance, Li *et al.* [27] adapted the original Shapley value computation, a classic perturbation-based approach, to lane-change models, calculating attributions based on marginal contributions and significantly enhancing the decision interpretability. Sacha *et al.* [28] leveraged semantic segmentation information to adaptively identify key features within the DNN. DeepLIFT, introduced by Shrikumar *et al.* [29], compares neuron activations against a reference state to trace the influence of different neurons. While versatile, perturbation-based methods often suffer from computational inefficiency due to the repeated iterations or operations required for each input perturbation.

Propagation-based methods leverage back-propagated gradients and forward-propagated feature representations for attribution computation. Chormai *et al.* [30] enhanced layer-wise relevance propagation (LRP) by integrating principal component analysis and independent component analysis to extract concept subspaces, focusing attributions on information actively used by the DNN. Yang *et al.* [31] introduced important direction gradient integration (IDGI), incorporating the concept of important directions into integrated gradients to effectively reduce noise in pixel-attribution results. Chen *et al.* [32] proposed PropShapley, a Shapley-value propagation model designed to facilitate attribution computation across different modalities within DNNs. Zhang *et al.* [33] developed salient manipulation path (SAMP), a path attribution method that efficiently identifies a near-optimal manipulation path from a predefined set, validating its effectiveness on several image classification datasets. Li *et al.* [34] analyzed the influence of model weights and sample features on LRP, leading to the development of weight-dependent baseline LRP (WB-LRP) for graph convolutional neural networks. GradShap [35] has been successfully applied to various autonomous driving scenarios, including traffic object detection [36] and lane-change prediction [37], effectively revealing the rationale behind specific decisions. SIAS [38] enhanced the interpretability of attribution explanations by incorporating semantic labels. This addition provides a more human-understandable context for the attributions, linking the model's decisions to recognizable semantic blobs. Furthermore, several DNN attribution methods originally designed for fields like biology, economics, and psychology [39]–[42] could be adapted, with minor modifications, to enhance the explainability of autonomous driving models.

Other attribution methods combine techniques like propagation, perturbation, and attribution baselines, as seen in approaches such as LRP optimization [43] and attribution aggregation [44]. Although existing methods effectively establish input-output causality, many are highly specialized and challenging to adapt or generalize to different autonomous driving models without sacrificing accuracy. This lack of

generalizability results in inconsistent attribution computation principles across various models, directly impacting the reliability and comparability of the generated explanations.

B. Explainable Autonomous Driving Models

The use of attention mechanisms for building interpretable autonomous driving models has been extensively explored across various tasks, including object detection [45], [46], motion forecasting [47], [48], driver attention prediction [49], and recent end-to-end models [50]–[52]. Attention mechanisms can highlight important features, offering insights into the model's decision-making process. Moreover, they only introduce minimal computational overhead, facilitating real-time interpretability analysis.

However, compared to attribution methods, attention mechanisms face two key challenges in practical applications. First, attribution methods can benefit from theoretical guarantees, such as the axiom constraints inherent in Aumann-Shapley attributions, providing a stronger theoretical foundation for the explanations. Attention computations, on the other hand, often lack such formal interpretability constraints. Second, attention models typically exhibit architectural specificity, performing optimally within self-attention-based networks and potentially underperforming or being inapplicable in other architectures.

III. Preliminary of Aumann-Shapley Attribution

To address the opacity of the decision-making process in DNN models, the use of Aumann-Shapley (AS) attribution has proven effective and offers many desirable properties. This method, rooted in cooperative game theory, measures the contribution of individual players to a specific outcome. When applied to deep learning, it provides a principled way to determine how input features contribute to a prediction.

For a decision-making model f used in autonomous driving, given the input data $x^{[0]}$, it is processed by the model to produce the output $f(x^{[0]})$. Furthermore, the feature maps obtained at the middle layer l can be denoted as $x^{[l]} = f^{[l]}(x^{[0]})$, where the superscript $[l]$ for x is omitted for simplicity in subsequent contents. Consequently, $f(x)$ represents the network decision when the network is truncated at layer l, with x as the neuron features at this layer. In addition, we define the missing information state of x as $\bar{x}$, and describe the integration path from $\bar{x}$ to x as $\mu(t)$, where t ranges from 0 to 1, starting at $\mu(0) = \bar{x}$ and culminating at $\mu(1) = x$. The AS attribution evaluates the marginal contribution of each neuron i to the final decision, calculated as follows:

$$\phi_i = \int_{t=0}^{1} \left(f\left(\mu(t) + \frac{\partial \mu(t)_i}{\partial t}\right) - f\big(\mu(t)\big) \right) dt, \quad (1)$$

where $\partial \mu(t)_i / \partial t$ quantifies the infinitesimal change in the i-th feature x_i at t. Here, i ranges from 1 to N, where N is the total number of features. The path $\mu(t)$ has the same dimensionality as x. The integral calculates the cumulative effect of this change on the model's output, with ϕ_i quantifying the marginal contribution of x_i to the decision-making process.

We next expand the first term of the integrand using the Taylor series:

$$\begin{aligned} &f\left(\mu(t) + \frac{\partial \mu(t)_i}{\partial t}\right) \\ &= f\big(\mu(t)\big) + \frac{\partial f\left(\mu(t)\right)}{\partial \mu(t)_i} \frac{\partial \mu(t)_i}{\partial t} + O(dt^2), \end{aligned} \quad (2)$$

where the remainder term $O(\cdot)$ accounts for higher-order corrections which are negligible as dt approaches zero. Thus, Eq. (1) simplifies to:

$$\phi_i = \int_{t=0}^{1} \frac{\partial f\big(\mu(t)\big)}{\partial \mu(t)_i} \frac{\partial \mu(t)_i}{\partial t} dt, \quad (3)$$

where the first term of multiplier represents the gradient of the prediction with respect to the i-th neuron feature, and the second term denotes the path direction of the i-th feature. The resulting attributions adhere to several foundational axioms of attribution explanation, ensuring that the results are theoretically sound and meaningful. The core axiom is "efficiency," which states that the sum of the attributions, $\sum_i \phi_i$, must equal the difference in the model's output between the original input and the starting point: $f(\mu(1)) - f(\mu(0))$. This axiom serves as a sanity check for the generated attribution scores. Further discussion regarding attribution axioms can be found in [53].

IV. Traffic Scene-Informed Attribution Computation

Although AS attribution offers ideal theoretical properties, its direct application to autonomous driving decision-making models presents several challenges. These arise primarily from the diverse nature of model tasks and the complexity of traffic scenarios. For example, when applied to interpreting decisions made by lane-change models, the attributions can often be incorrect or misleading. Moreover, even in the context of vision-based models, the attribution results may appear counterintuitive and are frequently affected by noise. These issues motivate us to think about how to develop a unified attribution computational model to improve the quality of attribution in the face of multiple autonomous driving tasks.

A key source of inaccurate attributions stems from arbitrarily chosen integration paths, such as straight-line paths, which often fail to incorporate crucial scene information. Due to the dispersed and complex nature of relevant data in traffic scenarios, such paths can inadvertently traverse regions with high-magnitude gradients and feature maps corresponding to irrelevant information, leading to misleading attributions that misrepresent the model true decision-making process. Next, we will outline the method for designing a path that is informed by traffic scenes, as well as determining its optimal starting point. Fig. 2 provides a concise overview of our attribution computation pipeline.

To reduce misleading and noisy attributions, we incorporate traffic scene information to guide the integration path direction for our attribution model. We specifically utilize the first-order term of the Taylor expansion of a DNN, *i.e.*, the product of the gradient and feature map (GFM) to extract contextual insights. Here, acknowledging that changes in the resolution of the feature map initiate feature aggregation and abstraction

processes, we leverage GFM from the subsequent layer/block (denoted by $l+1$) when computing attributions for the layer l.

The distance d between the GFM of a point on the path and the target GFM serves as the objective function of path generation. As the integration parameter t increases, the traffic information represented by the corresponding point on the path progressively transitions from a state of missing information to a complete representation of the current traffic scene. Thus, the objective function d can be defined as:

$$d\big(\mu(t)\big) = \left\| \frac{\partial f\big(\mu(t)\big)}{\partial f^{[l+1]}\big(\mu(t)\big)} f^{[l+1]}\big(\mu(t)\big) - \frac{\partial f(x)}{\partial f^{[l+1]}(x)} f^{[l+1]}(x) \right\|_2, \tag{4}$$

where d represents the L2 distance used to evaluate the discrepancy between the GFM of the current state along the path $\mu(t)$ and the GFM of the original input state. The size of $\mu(t)$ matches the input x, while the size of $f^{[l+1]}(\cdot)$ corresponds to the feature map $x^{[l+1]}$ at layer $[l+1]$. This objective function aims to ensure that the points along the path rapidly minimize the difference between the current GFM and the target GFM. This process effectively represents a progressive approach towards the target traffic scene. In essence, the function guides the path along the steepest descent direction within the GFM space, promoting the fastest convergence to the desired traffic scenario.

Since DNN feature maps are discrete, the path connecting two GFMs is represented by a sequence of sampled points. Generating this path can be reframed as optimizing the generation of these discrete sample points. Then, the path generation becomes an optimization problem: finding the optimal direction $p(m)$ and step size $\eta(m)$ for each iteration m to minimize the objective distance. The direction $p(m)$ can be directly derived from the distance objective as follows:

$$p(m) = \frac{\partial d / \partial x'}{\|\partial d / \partial x'\|_2}, \tag{5}$$

where $x' = \mu(m/n)$ represents the intermediate state along the path at the given step m/n, within the discrete set defined as $\{0/n, ..., m/n, ..., n/n\}$. Here, p is the normalized gradient of the distance d, which provides the direction that most rapidly decreases the distance. Given $\mu(0)$, $\mu(1)$, and the direction of each step $p(m)$, we then determine the step size $\eta(m)$ by optimizing a simple objective function:

$$\arg\min_{\eta(m)} \left\| \mu(1) - \big(\mu(0) + \sum_m p(m)\eta(m)\big) \right\|_2 \tag{6}$$

where $\eta(m)$ represents the step size. The optimized p and η allow for precise control in regions with significant feature activity, while also enhancing efficiency across less critical areas, thereby maintaining the model's accuracy and computational viability. As a result, the traffic scene-informed path is updated by

$$\mu\left(\frac{m+1}{n}\right) = \mu\left(\frac{m}{n}\right) + p(m)\eta(m). \tag{7}$$

Having established the methodology for calculating the path, it is crucial to also determine another key parameter

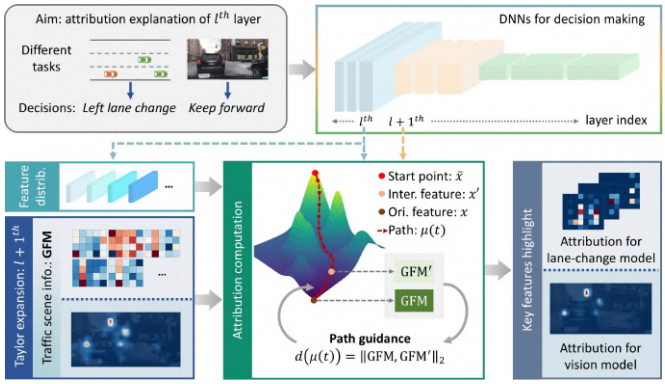


Fig. 2. Brief pipeline of our attribution computation. Our method accommodates different types of autonomous driving models without requiring any modifications. By leveraging traffic scene information and feature distribution, we define the complete attribution calculation path, ultimately generating attribution explanations in different models.

that significantly influences the attribution results: the starting point $\bar{x}$ of the path (*i.e.*, $\mu(0)$). This point is also considered as an attribution baseline representing the lack of information with respect to x and is chosen to trigger decision shifts.

In prior applications of AS attribution, especially with image classification models, the starting point typically aims to reduce the model's prediction score to zero, representing the absence of the target information. However, we find this assumption inappropriate for autonomous driving models. Unlike classification scenarios with thousands of labels, lane-change models typically have only three decisions (keep lane, left lane change, right lane change), and end-to-end vision models often have only four (keep forward, stop, turn left, turn right).

Consequently, changing a model's decision only requires lowering the current decision's score below the average. This signifies a lack of information supporting the original decision. For instance, in a three-output lane-change model, reducing the decision score below $s = 1/3$ during starting point optimization fulfills the core principle of AS attribution. With three output categories and a total probability of 1, any score below average necessitates a decision change.

To enhance optimization, we further reduce this threshold by 10% to 0.3. This ensures a robust decision change and a higher-quality starting point. The score loss is then defined accordingly:

$$\mathcal{L}_{\text{score}} = \big(f(\bar{x}) - s\big)^2, \tag{8}$$

where $\bar{x}$ is the starting point, s stands for the score target. This loss can reduce the score just enough to trigger a change in decision.

Furthermore, the attribution computation model in Eq. (3) relies on omitting higher-order terms in the Taylor expansion. This approximation requires that the starting point $\mu(0)$ and endpoint $\mu(1)$ remain close. We design this constraint based on the feature distribution observed in autonomous driving scenarios. First, we estimate the hidden layer feature distribution using the entire dataset. As the spatial dimensions of feature maps generally lack universality, we retain only the 10th percentile of each channel in the feature maps as

constraints. Because a small percentile typically represents the lower bound of a specific feature's distribution, we use this value to constrain the optimization range of $\bar{x}$:

$$\begin{aligned} \mathcal{L}_{\text{range}} &= \max\left(\||x-\bar{x}|-\tau\|_2, 0\right), \\ \tau &= \mathbb{E}_{x^{[0]}\sim\mathcal{X}}\left[Q_{0.1}\left(f^{[l]}\left(x^{[0]}\right)\right)\right], \end{aligned} \tag{9}$$

where $x^{[0]}$ represents an initial input sampled from the input distribution $\mathcal{X}$, and $f^{[l]}(\cdot)$ denotes the output after the layer l. $Q_{0.1}$ means the 10th percentile value, calculated independently for each channel. Then, the overall optimization objective for the starting point $\bar{x}$ is:

$$\arg\min_{\bar{x}} \mathcal{L}_{\text{score}} + \lambda \mathcal{L}_{\text{range}}, \tag{10}$$

where λ is used to balance two terms. We directly utilize the Adamax optimizer during the optimization process. However, in addition to the score and range constraints, $\bar{x}$ has another special property. Because it represents the absence of information relevant to the current decision, the gradient produced by a trained model for an input representing missing features should have a larger magnitude than that for a common input. This aligns with the principles of model training, where outliers typically generate larger fluctuations in gradient updates. Therefore, we introduce a hard constraint, *i.e.*, the absolute value of the gradient at $\bar{x}$ must be bigger than the absolute value of the gradient at the original input x:

$$\left|\frac{\partial f(x)}{\partial x_i}\right| \leq \left|\frac{\partial f(\bar{x})}{\partial \bar{x}_i}\right|. \tag{11}$$

If, during the optimization process, a specific neuron violates the constraint, we apply a small step update until it returns to the valid range: $\bar{x}_i \leftarrow \bar{x}_i + \beta \text{sgn}\left(\partial f(\bar{x})/\partial \bar{x}_i\right)$. We set the step size, β, to 70% of the optimizer's current learning rate. Although a smaller β could theoretically improve precision, it would also negatively impact optimization efficiency. Conversely, a value close to the learning rate does not introduce noticeable numerical errors. With the complete path defined, we can compute the attributions for any layer of a given autonomous driving model.

V. Experiments

A. Datasets and Models

Autonomous driving datasets. Our experiments primarily utilize two autonomous driving datasets: HighD [54] and BDD [55], [56].

The HighD dataset focuses on highway lane-change maneuvers collected using drones on German highways. It encompasses over 110,500 vehicles and encompassing a variety of lane-change scenarios, including left and right lane changes, as well as forced lane changes. A key feature of the dataset is the high-frequency data acquisition at 25Hz, allowing for detailed analysis of subtle changes in vehicle behavior during these maneuvers. Furthermore, HighD offers rich information on vehicle interactions, including the speed, acceleration, and distances of surrounding vehicles, providing crucial context for understanding the complex dynamics involved in lane changes.

The BDD dataset consists of 10,000 video clips (totaling approximately 1,000 hours) primarily collected in New York, Berkeley, San Francisco, and the Bay Area. It includes synchronized GPS/IMU data from mobile devices, enabling approximate trajectory reconstruction, and spans a wide range of weather conditions (sunny, overcast, rainy) and lighting scenarios (daytime/nighttime). BDD also provides multi-task annotations for perception tasks, including object detection, semantic segmentation, lane detection, and instance segmentation, making it a valuable resource for training and evaluating autonomous driving models.

Autonomous driving models. Our experiments mainly include two models: a discretionary lane-change (DLC) model [19] based on contextual information and the driving styles of surrounding vehicles, and an end-to-end vision-based autonomous driving model.

The DLC model [19] leverages the Driving Operational Picture (DOP) as a key feature representing driving style. The DOP information is input into a DNN to predict one of three possible decisions: lane keep, left lane change, or right lane change. DOP is structured as a matrix comprised of seven statistical features and eight vehicle features. The rows of the matrix represent the statistical measures: mean, standard deviation, median, 25th percentile, 75th percentile, minimum, and maximum, calculated across the eight vehicle features. These vehicle features, forming the columns of the matrix, include relative longitudinal position (X), relative lateral position (Y), longitudinal velocity, lateral velocity, longitudinal acceleration, lateral acceleration, space headway, and time headway. The vehicles considered for feature extraction are the ego vehicle (Ego), preceding vehicle (P), preceding vehicle in the left and right adjacent lane (LP and RP), following vehicle in the left and right adjacent lane (LF and RF), and alongside vehicle in the left and right adjacent lane (L and R). That is, the input of the DLC model includes eight DOP matrices. The model incorporates a two-second reaction time assumption for lane-change decisions, reflecting typical human driver behavior. Training of the DLC model is performed using the HighD dataset.

The vision-based model employs a DenseNet architecture [25], adapted for autonomous driving, and trained on the BDD dataset [55], [56] following the implementations described in [2], [57]. This densely connected network takes a visual image as input and predicts one of four driving actions: keep forward, brake, turn left, or turn right. The model training scheme can be categorized as imitation learning, because the training data labels are derived from image annotations provided by experienced human drivers.

B. Implementation Details

Parameter selection. In practice, our computation of attributions uses a discrete integration method with 100 sampling points ($n = 100$). This configuration ensures that the generated attributions closely approximate the output score, effectively satisfying the "efficiency" axiom. According to this axiom, the sum of the attribution values should closely match the model's output logit score. Satisfying the axiom not only enhances the credibility of the attributions but also facilitates their numerical validation. Increasing the number of sampling points reduces

the step sizes, theoretically improving accuracy. However, we observe diminishing returns beyond a certain point, with no significant gains in accuracy. Conversely, using too few sampling points increases the step size, often introducing uncontrollable noise.

Our empirical findings suggest that optimal results are achieved through a two-stage process: first, independently optimizing each term in Eq. (10) while monitoring their magnitude, then setting λ such that the value of $\mathcal{L}_{\text{range}}$ approximates half the value of $\mathcal{L}_{\text{score}}$. Excessively small λ value risks neglecting the constraint on $\bar{x}$, potentially generating adversarial samples, while overly large λ value impedes the convergence of $\mathcal{L}_{\text{score}}$. This parameter requires manual finetuning tailored to the specific model architecture and optimization algorithm.

Computational cost. Our experiments are conducted on two platforms: a server with dual NVIDIA RTX A6000 GPUs and a local machine with Intel I9 13900K CPU and an NVIDIA RTX 4090 GPU. The computational cost was performed on the local machine. Our attribution method, designed for offline explanation generation due to its iterative optimization workflow, requires 6.72 seconds to process the DLC model and 28.64 seconds for the vision model per input sample. The computation comprises two iterative phases: 1) starting point optimization dominates 64% of the total runtime, and 2) path optimization, which refines the integration path and accounts for 29% of the runtime. Even with batch processing and a batch size of 32, processing a single input sample for the vision model still takes approximately 0.9 seconds. These results position our proposal as a practical tool for post-hoc analysis rather than real-time onboard deployment in autonomous driving systems.

C. Attribution Comparison on Lane Change Model

In this section, we conduct attribution explanation experiments on the DLC model [19]. We begin with a quantitative comparison against other attribution methods adaptable to lane-change models, demonstrating the advantages of our proposal. Then, we proceed with a qualitative analysis, showcasing the explanatory power of our method through a representative attribution example.

Quantitative comparison. To quantitatively compare our approach, we compare it against several recent and relatively general attribution methods applied to the DLC model: PRCA [30], WB-LRP [34], IDGI [31], SVCE [27], PropShap [32], and SAMP [33]. Both PRCA and WB-LRP rely on layer-wise relevance propagation. Adapting their relevance propagation rules to specific hidden layer structures can be challenging. However, because the DLC model's hidden layers primarily consist of linear and ReLU activations, no specific rule adjustments are required for these layers. We handle the DLC model's multiple branches by distributing attributions proportionally to branch weights. IDGI, SVCE, PropShap, and SAMP are model-agnostic, requiring minimal modification for application to the DLC model. SVCE, originally applied to a reinforcement learning lane-change model using Shapley values, is adapted to the DLC model with random sampling to reduce computational complexity. Although attribution explanation generation is an offline process, we aim to maintain a reasonable per-sample explanation time (on the order of seconds to tens of seconds) to avoid excessive computational overhead.

TABLE I
QUANTITATIVE ATTRIBUTION COMPARISON ON THE DLC MODEL.

	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
PRCA	0.623	0.628	0.641	0.803	0.179
WB-LRP	0.551	0.597	0.613	0.739	0.199
IDGI	0.561	0.592	0.614	0.747	0.187
SVCE	0.512	0.581	0.607	0.716	0.192
PropShap	0.627	0.662	0.678	0.828	0.167
SAMP	0.642	0.657	0.664	0.831	0.152
Zero	0.626	0.665	0.685	0.821	0.166
Uniform	0.581	0.613	0.622	0.794	0.174
Gaussian	0.605	0.602	0.628	0.781	0.162
$\mathcal{L}_{\text{score}}$	0.609	0.609	0.614	0.775	0.147
DirectPath	0.608	0.649	0.671	0.808	0.171
GradPath	0.596	0.612	0.629	0.774	0.177
Ours	**0.685**	**0.717**	**0.723**	**0.889**	**0.134**

We use several quantitative metrics for comparison: Sensitivity-n (Sen-n) [53], [58], Accuracy Information Curve (AIC), Softmax Information Curve (SIC) [44], Least-Relevant-First (LeRF), and Most-Relevant-First (MoRF) [58]. Sen-n is a key indicator of the attribution "efficiency" axiom, crucial for theoretical soundness. AIC, SIC, LeRF, and MoRF share a common goal: quantifying how accurately attributions identify truly important features. AIC and SIC approach this evaluation from an information entropy perspective, while LeRF and MoRF directly utilize decision scores for a more direct assessment. Together, these metrics provide a comprehensive view of how accurately the attributions pinpoint critical features, thereby assessing the reliability and validity of the numerical attribution values.

The quantitative results (Table I) reveal several key findings. Shapley-based methods generally perform better in terms of Sen-n, demonstrating a stronger correlation between the impact of attributed features and their assigned attribution values. However, even our best-performing method achieves a Sen-n score below 0.7 for the DLC model. This suggests that maintaining high sensitivity, despite theoretical guarantees, becomes increasingly difficult in practice with greater model complexity. AIC and SIC evaluate the ability to progressively recover features based on their attributions, indirectly reflecting the accuracy of capturing critical feature contributions. LeRF and MoRF offer a more direct assessment of the attribution distribution's validity. Our method excels across these metrics, notably achieving a LeRF score approaching 0.89. This indicates our attributions accurately identify and appropriately weight critical information.

Ablation study. Our method primarily modifies the integration path used in attribution calculation, which involves selecting the path's starting point and generating the path itself. The starting point acts as the baseline for attribution calculation. Existing research [53] has explored various baseline types for attribution, including blurred features, random noise,

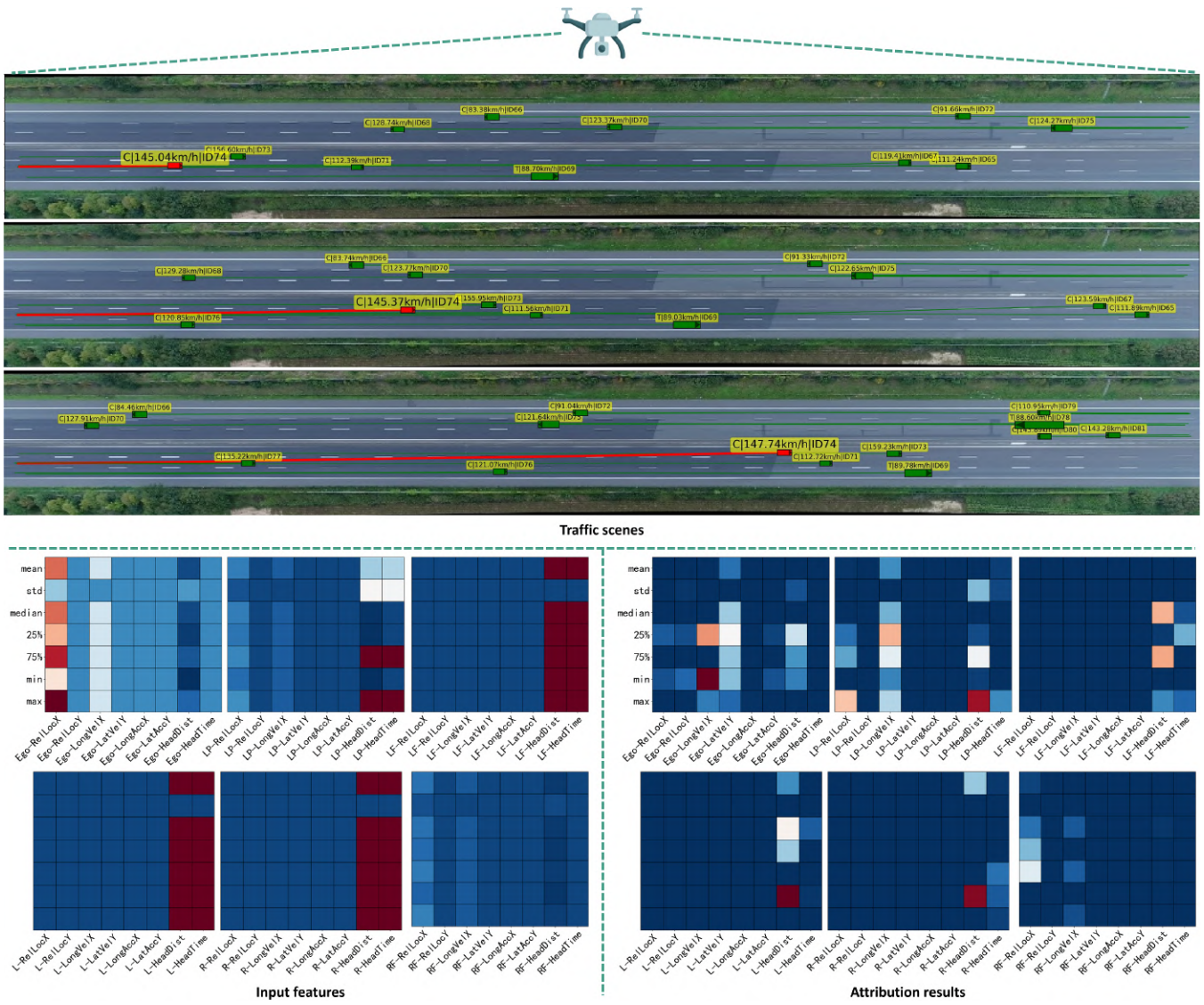


Fig. 3. Lane-change model attribution explanation. The upper part displays the driving scene where the lane change occurs, with the red vehicle being the maneuvering vehicle. Yellow labels above each vehicle indicate their respective speeds and IDs. The lower-left part shows the input information to the model. The lower-right part presents the corresponding attribution results, visualizing the contributions of different input features to the model's lane-change prediction.

and the expected value of network features. However, many existing baselines are often specifically designed for vision models and may not be directly applicable to other domains. For autonomous driving applications, we adopt three directly applicable baselines: zero, uniform noise, and Gaussian noise. We conduct comparative experiments using these baselines, with uniform noise bounds and Gaussian parameters (mean and variance) derived from the dataset's feature distribution.

A critical requirement for baselines is their ability to represent feature absence in decision-making. For autonomous driving models, any baseline that induces a measurable decision shift satisfies this criterion. Building on this principle, we design an additional baseline using only the $\mathcal{L}_{\text{score}}$ term from Eq. (10) for ablation studies. Our ablation study evaluates four baseline variants for the attribution path starting point: 1) Zero, 2) Uniform, 3) Gaussian, and 4) $\mathcal{L}_{\text{score}}$. As shown in Table I, all baselines yield suboptimal attribution results. Notably, the zero baseline outperforms other alternatives, suggesting that simpler baselines may better align with the feature absence criterion than introducing noise or $\mathcal{L}_{\text{score}}$ baselines.

Furthermore, we design two ablation experiments regarding the integration path itself. These involve: 1) DirectPath: Replacing our proposed path with a direct linear path; 2) GradPath: Using gradients instead of GFM for guidance. The results show that GradPath performs worst, suggesting that gradient information may not effectively capture the implicit scene representation. The DirectPath also exhibits obvious decreased performance compared to our baseline.

Qualitative analysis. Fig. 3 depicts a lane-change scenario involving the red vehicle (ID74). The upper part of the figure illustrates the overall traffic situation during the lane-change maneuver, captured by drone footage over a specific road segment. The lower-left part presents a subset of the model's input information, while the lower-right displays the corre-

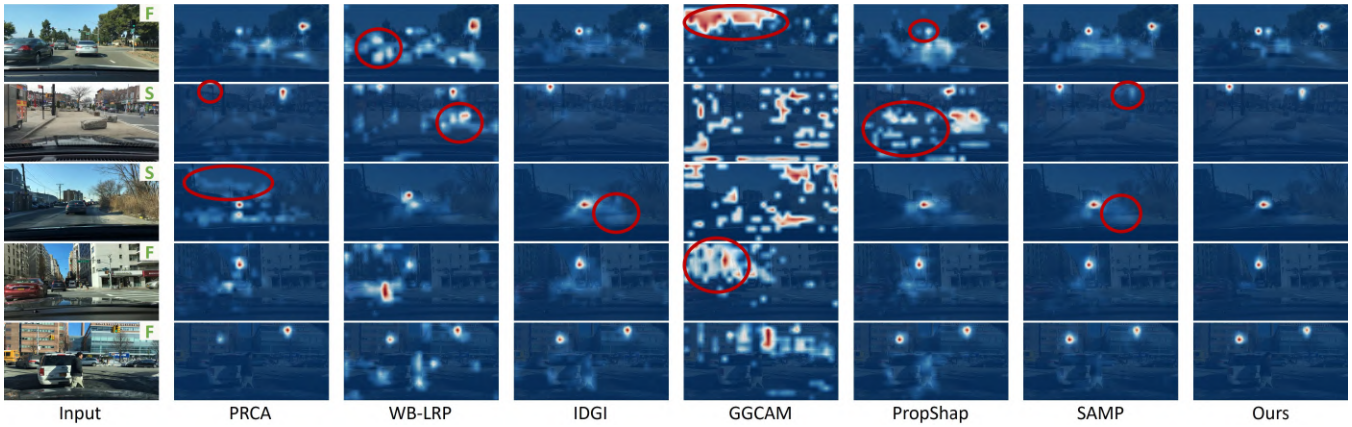


Fig. 4. Attribution explanations on vision model. The green characters in the top-right corner of the original images indicate the model's decision (F: Forward, S: Stop). Red circles highlight areas of attribution noise and potentially misleading artifacts. In these examples, traffic lights always significantly influence the model's output. Because of their small size within the images, these figures are best viewed on screen for optimal clarity.

sponding attributions generated by our method. For clarity, only the inputs and attributions with significant influence on the output are shown. Absent vehicles (LF, L, and R in this case) are also represented in the input matrices. A deep red color in the head distance and head time signifies infinite distance for these absent vehicles. Their corresponding X and Y positions, along with other relevant information, are set to zero, visually represented by a deep blue. Because the predicted lane change utilizes information from before the maneuver occurs, the input features in the lower-left part do not directly correspond to the traffic scenes depicted above. The traffic scene information primarily serves to provide context for understanding the lane-change situation.

Given the input, the model correctly predicts a left lane change. However, the DNN's decision-making process remains opaque. Our generated attribution explanation reveals the key factors influencing this prediction. The most significant contributors are: 1) the Ego vehicle's minimum and 25th percentile longitudinal velocities; 2) the LP vehicle's longitudinal velocity and its head distance (distance to the vehicle in front, often infinite due to the LP frequently having no preceding vehicle); 3) the absence of LF, L, R vehicles; and 4) the relative position of the RF vehicle. Although all input features receive attributions; for clarity, we only list these six input feature matrices that significantly influence the model prediction.

This result suggests that the model's processing of driving style features aligns, to some extent, with human driving intuition. The surrounding context shows the ego vehicle approaching the lead vehicle at a relatively high speed without decelerating, implying an intention to change lane, as otherwise, a collision with the preceding vehicle would be imminent. A safe left lane change requires assessing the left lane's state, including the distance and speed of any left-following vehicle and the presence of any vehicles alongside. The speed of the preceding vehicle in the target lane is also crucial for a safe maneuver. Simultaneously, the conditions behind the ego vehicle must be considered. These considerations align with the highlighted attributions, demonstrating that the driving style-enhanced DLC model effectively leverages driving style information for its predictions. Attribution explanations, therefore, not only illuminate input-output relationships but also reveal parallels between model decisions and human cognition, providing valuable insights into the behavior of autonomous driving models.

TABLE II
QUANTITATIVE ATTRIBUTION COMPARISON ON THE VISION MODEL.

	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
PRCA	0.739	0.622	0.641	0.767	0.162
WB-LRP	0.719	0.589	0.602	0.748	0.189
IDGI	0.741	0.615	0.622	0.762	0.159
GGCAM	0.604	0.552	0.568	0.701	0.251
PropShap	0.734	0.624	0.647	0.778	0.165
SAMP	0.779	0.658	0.676	0.794	**0.125**
Zero	0.747	0.613	0.625	0.771	0.158
Uniform	0.707	0.607	0.623	0.739	0.182
Gaussian	0.732	0.622	0.636	0.778	0.147
$\mathcal{L}_{\text{score}}$	0.711	0.627	0.633	0.759	0.139
DirectPath	0.722	0.607	0.619	0.762	0.166
GradPath	0.716	0.582	0.595	0.723	0.173
Ours	**0.822**	**0.683**	**0.711**	**0.835**	0.136

D. Attribution Comparison on Vision Model

In this section, we apply our attribution method to an end-to-end vision model, evaluating its performance through both qualitative and quantitative experiments. For comparison, we also incorporate a common attribution method, Guided Grad-CAM, fine-tuned based on established techniques [50], [59]. Unlike the lane-change model which uses structured data, the vision model receives an image as input. Information is progressively downsampled within the model, and features become more abstract with decreasing pixel resolution. Although individual pixels or neurons lack inherent semantic meaning, feature attributions at any hidden layer can be upscaled to the original image size, creating a heatmap that highlights important image regions for interpretability. We select a hidden layer with a three-fold reduction in image size for attribution computation. This choice is supported by previous research indicating that such high-level features exhibit object corre-

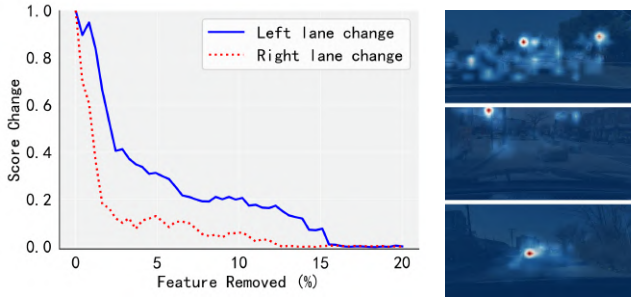


Fig. 5. Analysis of GFM's representation ability. a) Changes in the DLC model's prediction score after removing features based on GFM; b) GFM heatmaps that can roughly highlight key regions for the output of the vision model.

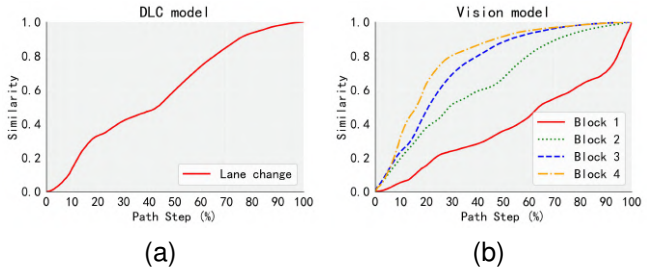


Fig. 6. GFM similarity along attribution path. a) Similarity results for the DLC model; b) Similarity results for the four main hidden layers of the vision model.

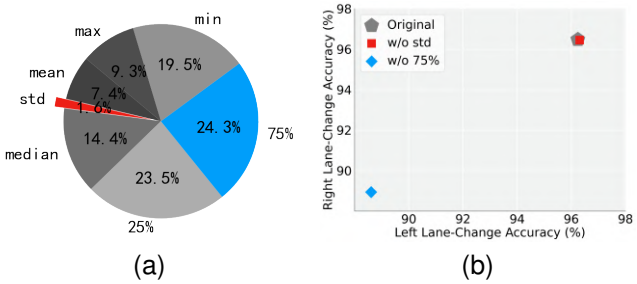


Fig. 7. Impact of different statistical features on the model. a) Attribution contribution of different features; b) Accuracy of different lane-changing models.

spondence [60], allowing attributions to be mapped to specific objects for easier analysis.

Fig. 4 provides a qualitative comparison. It demonstrates that attribution computations without guidance from traffic scene information can be misleading (highlighted by red circles). Masking these regions (replacing them with a gray background) does not significantly decrease the output score and, in some cases, even increases it. We also use quantitative metrics for comparison, and as shown in Table II, our method achieves the best performance across most metrics.

E. Analysis of Traffic Scene Representation

In this section, we analyze GFM's ability to represent traffic scene information across different autonomous driving tasks. For the DLC model, which uses statistical features as input, the corresponding GFM is an abstract representation of these features. We design an intuitive evaluation metric to assess this representation. Features are sorted according to their GFM values, and then progressively set to zero or infinity based on the scenario. The resulting change in the output score is then plotted. An effective representation should exhibit a significant downward trend in this plot, as removing relevant information should decrease the decision score. As shown in Fig. 5a, both left and right lane-change decisions demonstrate this downward trend for the DLC model. The decision score drops to zero after removing approximately 20% of the features, demonstrating GFM's ability to represent the scene.

For the vision model, we directly visualize the GFM as a heatmap (Fig. 5b). The resulting heatmap generally corresponds to the key regions influencing the decision, indirectly demonstrating GFM's capacity to represent different types of traffic scene information.

Experiment results confirm that the first-order term of the Taylor expansion at a specific hidden layer effectively captures decision-relevant traffic scene information, validating its use for guiding the integration path in attribution computation.

F. Analysis of Traffic Scene Reconstruction

In this section, we analyze the progressive reconstruction of traffic information along our attribution path. Our attribution computation model's integration path is guided by GFM. The intermediate states, x', along this path represent the gradual reconstruction of traffic information. To analyze this process, we use the GFM corresponding to the path's endpoint (*i.e.*, x) as a reference. We then calculate the cosine similarity between the GFM corresponding to x' and the reference GFM to analyze how the traffic scene information is reconstructed along the path.

As shown in Fig. 6a, for the DLC model, we directly plot the GFM information from the subsequent layer used in the attribution calculation for the input. For ease of comparison, we normalize the different similarity measures to the range [0, 1], calculate the average values, and present the results in Fig. 6. The similarity gradually increases along the path, demonstrating a gradual reconstruction of the traffic scene. For the vision model, which has multiple hidden layers, we select four blocks where resolution changes occur for analysis. Fig. 6b shows that lower layers struggle to reconstruct scene features, while higher layers reconstruct them more easily. This observation aligns with the general understanding that lower-level features in DNNs typically represent basic, low-level visual elements, whereas higher-level features encapsulate more complex and abstract scene information. The greater detail present in low-level features contributes to the increased difficulty in their reconstruction. The reconstruction process exhibits a smooth and clear progression on these two different models and various hidden layers, indicating that our integration path effectively captures the gradual emergence of traffic scene information.

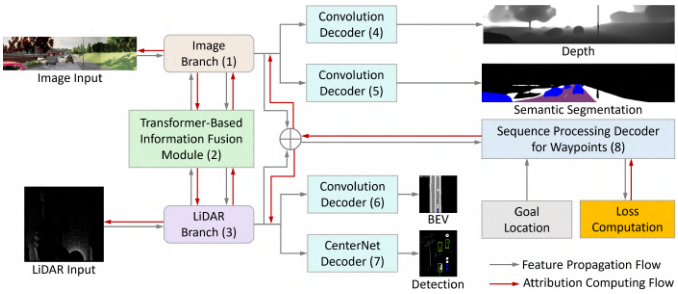


Fig. 8. TransFuser architecture. It comprises eight interconnected submodules designed to process multimodal sensor inputs (camera image and LiDAR), and to output trajectory-guiding waypoints for autonomous vehicles. The illustration was inspired by [52].

G. Attribution-based Feature Selection

In this section, we use attributions to improve feature selection for the DLC model. Our analysis of attribution explanations reveals a consistently low contribution from standard deviation information across predictions. To quantify this observation, we aggregate attribution values across all samples for each of the seven statistical metrics used as input features: mean, standard deviation, median, 25th percentile, 75th percentile, minimum, and maximum (Fig. 7a). The result reveals that the standard deviation contributes only 1.6% to the overall attributions across all lane-change predictions. Conversely, the 25th percentile, 75th percentile, and minimum exhibit substantially higher contributions, with the 75th percentile contributing the most at 24.3%.

Motivated by this finding, we conduct experiments where we ablate either the standard deviation or the 75th percentile feature during model training. The resulted models' accuracies are shown in Fig. 7b. Remarkably, removing the standard deviation feature does not negatively impact the model's predictive accuracy. In contrast, removing the 75th percentile feature results in a significant decrease in accuracy. This suggests that the standard deviation feature is redundant in the model's input space and can be removed to reduce computational overhead. The experiment also demonstrates the utility of attribution explanations not only for interpreting model outputs but also for informing the design and optimization of DNN models for autonomous driving, enabling more efficient model architectures.

H. Generalization Testing on Multimodal Model

To further validate the generalizability of our method, we conduct a qualitative attribution analysis using TransFuser [52], a state-of-the-art autonomous driving model known for its strong performance and accuracy on the CARLA urban driving benchmark [61]. TransFuser's Transformer-based architecture uses attention mechanisms to effectively fuse multimodal information from camera images and LiDAR data. As shown in Fig. 8, the model comprises eight interconnected submodules. The combination of self-attention, multimodal input, and complex architecture presents a challenging scenario for attribution methods. Successfully applying our approach to this intricate model demonstrates its robustness and adaptability to modern autonomous driving architectures.

Fig. 9 presents the attribution results, demonstrating our method's effectiveness in revealing the features influencing waypoint predictions. Comparing the image and LiDAR attributions highlights the distinct roles of each modality. For instance, LiDAR struggles to detect traffic lights (left and middle panels of Fig. 9), while the image branch effectively localizes them. During a sudden accident scenario (right panel of Fig. 9), where a bicycle collides with a gray car, the attribution reveals the model's focused attention on collision-relevant regions, demonstrating the model's ability to capture critical features even in complex and rapidly evolving situations. Notably, image attributions exhibit a dispersed pattern, suggesting the model leverages visual data for global scene understanding. In contrast, image attribution on purely vision-based models tends to be more concentrated.

While our method is directly applicable to the TransFuser model, we identify a critical challenge in assessing these attributions. Unlike simpler models with score-like outputs, advanced autonomous driving models like TransFuser produce complex, multidimensional outputs, such as sequences of waypoints or control commands. This complexity complicates the use of existing attribution evaluation metrics, which are typically designed for score outputs. Consequently, our current analysis focuses on qualitative assessment through visual inspection of attribution heatmaps. This highlights a critical need for new evaluation metrics specifically tailored to the multidimensional output spaces of modern autonomous driving models. Despite this challenge, we believe future work will bridge this gap, advancing both attribution methods and metrics to enhance trust and transparency in complex autonomous systems.

VI. Conclusion

In this work, we introduce a unified attribution method capable of providing accurate explanations for autonomous driving models across different task types. Drawing inspiration from research on Taylor expansions of DNNs, we use the first-order term (*i.e.*, GFM) of the Taylor expansion at a higher-level hidden layer to implicitly represent traffic scene information. The GFM then informs both the design of the entire integration path for attribution computation, resulting in a traffic-scene-informed attribution method. A key advantage of our approach is its broad applicability. Unlike other attribution methods that often require task-specific adaptations, our method can be consistently applied to different autonomous driving models, such as lane-change models and end-to-end vision models, without modification. Furthermore, our method achieves state-of-the-art performance across multiple quantitative metrics, demonstrating its robustness and accuracy.

Although our method demonstrates some strengths, several promising future directions exist for further development and improvement. Firstly, although certain desirable properties of Aumann-Shapley values hold theoretically for simple neural networks, verifying these properties becomes challenging in practice as model complexity increases. Developing axiomatic verification metrics to better ensure the trustworthiness of

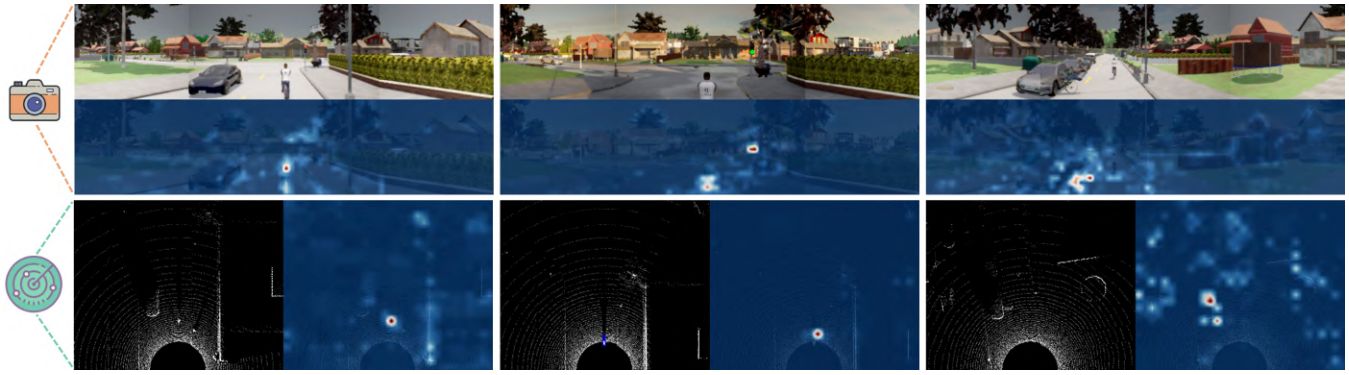


Fig. 9. Attribution results of the image and LiDAR branches. The attributions generated directly highlight the most important features impacting the TransFuser's waypoint prediction. These key features typically correspond to prominent objects within the image data and obstacles detected by the LiDAR sensor.

attributions in complex models is a priority for future research. Secondly, while our method proves effective across various autonomous driving models with different input modalities, evaluating attributions from diverse models, particularly multimodal ones, remains a crucial open problem. Developing robust evaluation metrics for these complex scenarios is essential for advancing the field. Thirdly, although Shapley values implicitly capture the contributions of feature combinations, current methods still evaluate these contributions through simple summation. A promising research direction involves analyzing the joint influence mechanisms of feature combinations on decision-making and designing interactive feature attribution methods that explicitly address these interactions.

REFERENCES

[1] G. Liu, J. Zhang, A. B. Chan, and J. H. Hsiao, "Human attention guided explainable artificial intelligence for computer vision models," *Neural Netw.*, vol. 177, p. 106392, 2024.

[2] Y. Feng, W. Hua, and Y. Sun, "NLE-DM: Natural-language explanations for decision making of autonomous driving based on semantic scene understanding," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9780–9791, 2023.

[3] W. Bao, B. Xu, and Z. Chen, "MonoFENet: Monocular 3D object detection with feature enhancement networks," *IEEE Trans. Image Process.*, vol. 29, pp. 2753–2765, 2020.

[4] J. Zhao, D. Wu, Z. Yu, and Z. Gao, "DRMNet: A multi-task detection model based on image processing for autonomous driving scenarios," *IEEE Trans. Veh. Technol.*, vol. 72, no. 12, pp. 15 341–15 355, 2023.

[5] Z. Shen, K. Cai, P. Zhao, and X. Luo, "An interactively motion-assisted network for multiple object tracking in complex traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1992–2004, 2024.

[6] B. Liang, W. Wei, J. Huang, C. Liu, H. Yang, R. Yang, W. Shang, and J. Li, "Real-time stereo image depth estimation network with group-wise L1 distance for edge devices towards autonomous driving," *IEEE Trans. Veh. Technol.*, vol. 72, no. 11, pp. 13 917–13 928, 2023.

[7] Y. Wang, D. Gloudemans, J. Ji, Z. N. Teoh, L. Liu, G. Zachár, W. Barbour, and D. Work, "Automatic vehicle trajectory data reconstruction at scale," *Transp. Res. Part C: Emerg. Technol.*, vol. 160, p. 104520, 2024.

[8] Y. Xue, C. Wang, C. Ding, B. Yu, and S. Cui, "Observer-based event-triggered adaptive platooning control for autonomous vehicles with motion uncertainties," *Transp. Res. Part C: Emerg. Technol.*, vol. 159, p. 104462, 2024.

[9] G. Du, Y. Zou, X. Zhang, Z. Li, and Q. Liu, "Hierarchical motion planning and tracking for autonomous vehicles using global heuristic based potential field and reinforcement learning based predictive control," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 8, pp. 8304–8323, 2023.

[10] S. Su, X. Ju, C. Xu, and Y. Dai, "Collaborative motion planning based on the improved ant colony algorithm for multiple autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2792–2802, 2024.

[11] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao, and H. Li, "Planning-oriented autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17 853–17 862.

[12] Y. Xiao, F. Codevilla, A. Gurram, O. Urfalioglu, and A. M. López, "Multimodal end-to-end autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 537–547, 2022.

[13] J. Townsend, T. Chaton, and J. M. Monteiro, "Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3456–3470, 2020.

[14] G. Liang, P. Tiwari, S. Nowaczyk, S. Byttner, and F. Alonso-Fernandez, "Dynamic causal explanation based diffusion-variational graph neural network for spatiotemporal forecasting," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2024.

[15] S. Rao, M. Böhle, and B. Schiele, "Better understanding differences in attribution methods via systematic evaluations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4090–4101, 2024.

[16] J. Hong, R. Hnatyshyn, E. A. D. Santos, R. Maciejewski, and T. Isenberg, "A survey of designs for combined 2D+3D visual representations," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 6, pp. 2888–2902, 2024.

[17] M. Li, H. Sun, Z. Cui, Y. Huang, and H. Chen, "Expected integral discrete gradient: Diagnosing autonomous driving model," *IEEE Trans. Veh. Technol.*, pp. 1–11, 2024.

[18] A. K. Gizzini, Y. Medjahdi, A. J. Ghandour, and L. Clavier, "Towards explainable AI for channel estimation in wireless communications," *IEEE Trans. Veh. Technol.*, vol. 73, no. 5, pp. 7389–7394, 2024.

[19] Y. Zhang, Q. Xu, J. Wang, K. Wu, Z. Zheng, and K. Lu, "A learning-based discretionary lane-change decision-making model with driving style awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 68–78, 2023.

[20] R. J. Aumann and L. S. Shapley, *Values of non-atomic games*. Princeton University Press, 1974.

[21] C. Wang, F. Guo, R. Yu, L. Wang, and Y. Zhang, "The application of driver models in the safety assessment of autonomous vehicles: Perspectives, insights, prospects," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 2364–2381, 2024.

[22] W. Liu, M. Hua, Z. Deng, Z. Meng, Y. Huang, C. Hu, S. Song, L. Gao, C. Liu, B. Shuai, A. Khajepour, L. Xiong, and X. Xia, "A systematic survey of control techniques and applications in connected and automated vehicles," *IEEE Internet Things J.*, vol. 10, no. 24, pp. 21 892–21 916, 2023.

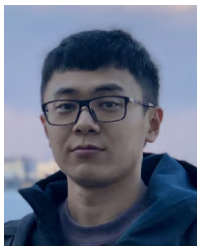
[23] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, 2017.

[24] H. Deng, N. Zou, M. Du, W. Chen, G. Feng, Z. Yang, Z. Li, and Q. Zhang, "Unifying fourteen post-hoc attribution methods with Taylor interactions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 4625–4640, 2024.

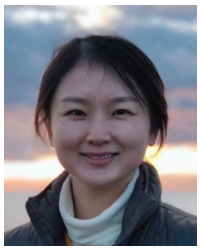
[25] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[26] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10 142–10 162, 2022.

[27] M. Li, H. Sun, Y. Huang, and H. Chen, "Svce: Shapley value guided counterfactual explanation for machine learning-based autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 14 905–14 916, 2024.
[28] M. Sacha, D. Rymarczyk, L. Struski, J. Tabor, and B. Zielinski, "ProtoSeg: Interpretable semantic segmentation with prototypical parts," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2023, pp. 1481–1492.
[29] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3145–3153.
[30] P. Chormai, J. Herrmann, K.-R. Müller, and G. Montavon, "Disentangled explanations of neural network predictions by finding relevant subspaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 11, pp. 7283–7299, 2024.
[31] R. Yang, B. Wang, and M. Bilgic, "IDGI: A framework to eliminate explanation noise from integrated gradients," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23 725–23 734.
[32] H. Chen, S. M. Lundberg, and S.-I. Lee, "Explaining a series of models by propagating Shapley values," *Nat. Commun.*, vol. 13, no. 1, p. 4512, 2022.
[33] B. Zhang, W. Zheng, J. Zhou, and J. Lu, "Path choice matters for clear attributions in path methods," in *Proc. Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=gzYgsZgwXa
[34] Y. Li, H. Liang, and L. Zheng, "WB-LRP: Layer-wise relevance propagation with weight-dependent baseline," *Pattern Recognit.*, vol. 158, p. 110956, 2025.
[35] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[36] A. Oseni, N. Moustafa, G. Creech, N. Sohrabi, A. Strelzoff, Z. Tari, and I. Linkov, "An explainable deep learning framework for resilient intrusion detection in IoT-enabled transportation networks," *IEEE Trans. Intell. Veh.*, vol. 24, no. 1, pp. 1000–1014, 2023.
[37] M. Li, Y. Wang, H. Sun, Z. Cui, Y. Huang, and H. Chen, "Explaining a machine-learning lane change model with maximum entropy Shapley values," *IEEE Trans. Intell. Veh.*, vol. 8, no. 6, pp. 3620–3628, 2023.
[38] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Understanding decision-making of autonomous driving via semantic attribution," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 1, pp. 283–294, 2025.
[39] H. Chen, I. C. Covert, S. M. Lundberg, and S. Lee, "Algorithms to estimate Shapley value feature attributions," *Nat. Mac. Intell.*, vol. 5, no. 6, pp. 590–601, 2023.
[40] D. Lundström, T. Huang, and M. Razaviyayn, "A rigorous study of integrated gradients method and extensions to internal neuron attributions," in *Proc. Int. Conf. Mach. Learn.*, vol. 162, 2022, pp. 14 485–14 508.
[41] J. D. Janizek, A. B. Dincer, S. Celik, H. Chen, W. Chen, K. Naxerova, and S.-I. Lee, "Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models," *Nat. Biomed. Eng.*, vol. 7, no. 6, pp. 811–829, 2023.
[42] C. Kim, S. U. Gadgil, A. J. DeGrave, J. A. Omiye, Z. R. Cai, R. Daneshjou, and S.-I. Lee, "Transparent medical image AI via an image–text foundation model grounded in medical literature," *Nat. Med.*, pp. 1–12, 2024.
[43] P. R. Bassi, S. S. Dertkigil, and A. Cavalli, "Improving deep neural network generalization and robustness to background bias via layer-wise relevance propagation optimization," *Nat. Commun.*, vol. 15, no. 1, p. 291, 2024.
[44] A. Kapishnikov, T. Bolukbasi, F. B. Viégas, and M. Terry, "XRAI: Better attributions through regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4947–4956.
[45] Z. Yuan, X. Song, L. Bai, Z. Wang, and W. Ouyang, "Temporal-channel transformer for 3D Lidar-based video object detection for autonomous driving," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2068–2078, 2022.
[46] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X. Hua, and M. Zhao, "Improving 3D object detection with channel-wise transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2723–2732.
[47] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2375–2384.
[48] B. Wei, M. Ren, W. Zeng, M. Liang, B. Yang, and R. Urtasun, "Perceive, attend, and drive: Learning spatial attention for safe self-driving," in *IEEE Int. Conf. Robot. Autom.*, 2021, pp. 4875–4881.
[49] C. Gou, Y. Zhou, and D. Li, "Driver attention prediction based on convolution and transformers," *J. Supercomput.*, vol. 78, no. 6, pp. 8268–8284, 2022.
[50] K. Ishihara, A. Kanervisto, J. Miura, and V. Hautamäki, "Multi-task learning with attention for end-to-end autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2902–2911.
[51] Q. Zhang, M. Tang, R. Geng, F. Chen, R. Xin, and L. Wang, "MMFN: Multi-modal-fusion-net for end-to-end driving," in *IEEE Int. Conf. Intell. Robot. Syst.*, 2022, pp. 8638–8643.
[52] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12 878–12 895, 2023.
[53] R. Shi, T. Li, and Y. Yamaguchi, "Output-targeted baseline for neuron attribution calculation," *Image Vis. Comput.*, vol. 124, p. 104516, 2022.
[54] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The HighD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2118–2125.
[55] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "BDD100K: A diverse driving dataset for heterogeneous multitask learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020.
[56] Y. Xu, X. Yang, L. Gong, H. Lin, T. Wu, Y. Li, and N. Vasconcelos, "Explainable object-induced action decision for autonomous vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9520–9529.
[57] X. Du, Z. Wang, M. Cai, and Y. Li, "VOS: Learning what you don't know by virtual outlier synthesis," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=TW7d65uYu5M
[58] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *Proc. Int. Conf. Learn. Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=S1xWh1rYwB
[59] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020.
[60] R. Shi, T. Li, L. Zhang, and Y. Yamaguchi, "Visualization comparison of vision transformers and convolutional neural networks," *IEEE Trans. Multim.*, vol. 26, pp. 2327–2339, 2024.
[61] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Annu. Conf. Robot Learn.*, 2017, pp. 1–16.



Rui Shi received his Ph.D. degree in graphic and computer sciences from the University of Tokyo, Tokyo, Japan, in 2022. He is currently a lecturer in the School of Information Science and Technology, Beijing University of Technology, Beijing, China. He worked as a visiting researcher in the Department of General Systems Studies, the University of Tokyo. His current research interests include autonomous driving, neural networks, and explainable artificial intelligence.



Tianxing Li received her Ph.D. degree in graphic and computer sciences from the University of Tokyo, Tokyo, Japan, in 2021. She is currently a lecturer in the College of Computer Science, Beijing University of Technology, Beijing, China. Her current research interests include neural networks and computer graphics.

Yasushi Yamaguchi received his Ph.D. in information engineering from the University of Tokyo in 1988. He is a professor of the Graduate School of Arts and Sciences, the University of Tokyo, Tokyo, Japan. His research interests lie in image processing, computer graphics, and visual illusion, including image editing, computer-aided geometric design, visual cryptography, and hybrid image. He was a former president of the International Society for Geometry and Graphics.

Liguo Zhang received his Ph.D. degree in control theory and applications from the Beijing University of Technology (BJUT), Beijing, China, in 2006. Since 2014, he has been a Full Professor with the School of Electronic Information and Control Engineering, BJUT. He is currently the Deputy Director of the School of Information Science and Technology, BJUT. His research interests include hybrid systems, intelligent systems, and control of distributed parameter systems. He is an Associate Editor for the IMA Journal Mathematical Control and Information and the Guest Editor of the International Journal of Distributed Sensor Networks.