Extending Aumann-Shapley for Reliable Neuron Attribution in Autonomous Driving Decisions

Rui Shi, Tianxing Li, Yasushi Yamaguchi, Liguo Zhang

***Abstract*—The rapid evolution of autonomous driving necessitates a clear understanding of the decision-making processes within these complex systems. While traditional path integral methods, which identify critical inputs by integrating gradients, offer a potential solution for attributing model predictions, their direct application to autonomous driving models often yields unreliable and counterintuitive results. This paper identifies two primary reasons for these shortcomings: the use of unreliable attribution baselines and integration paths. To address these problems, we propose a reliable method that constructs an open connected space for the baseline, specifically adapted to the unique attributes of driving scenes. Additionally, we implement baseline constraints to ensure the baseline accurately represents features absent in decision-making, thereby creating a solid foundation for accurate attribution computation. We also adjust the integration paths to accommodate the dispersed nature of objects in driving scenarios by using gradient-weighted feature maps, which helps in reducing noise and improving the reliability of attribution results. Extensive experimental results demonstrate the efficacy of our method, providing not only reliable and interpretable attributions but also outperforming state-of-the-art explanation techniques in both qualitative and quantitative evaluations.**

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

I. Introduction

A. Problem Formulation

AUTONOMOUS driving, empowered by artificial intelligence, has attracted widespread interest due to its potential to enhance convenience, improve safety, and offer economic benefits [1]–[5]. By employing deep neural networks (DNNs) with multimodal inputs, a range of applications including traffic object detection, dynamic path planning, scenario understanding, trajectory tracking, and decision-making have been significantly improved [6]–[12]. These advancements have significantly enhanced the ability of autonomous vehicles to navigate complex environments with greater precision and reliability, marking a paradigm shift in autonomous driving research.

Despite the critical role of DNNs in advancing autonomous driving technologies, they face a significant challenge due to their "black-box" nature, which makes their decision-making processes difficult to interpret. To address this, we consider the feasibility of using a generalized attribution method within the gradient integration framework, specifically the Aumann-Shapley (AS) [13] approach, known for its effectiveness and robustness among different complex models. This method quantifies the contributions of individual input features by measuring their incremental impacts as they are introduced sequentially from a baseline state of missing information to the full input scenario. However, when applying the AS approach to autonomous driving decision models, we observe that the attribution explanations often appear counterintuitive and lack credibility. As shown in the upper part of Fig. 1, the conventional AS method often generates attributions lacking cross-layer consistency, leading to unreliable and unintuitive explanations. Furthermore, based on the theoretical analysis of integrated gradient [14]–[16], we further discover that the primary cause of unreliable interpretations in autonomous driving scenarios can be traced back to two key techniques in the AS computational model.

Unreliable baseline. Baseline refers to a reference that represents the absence of features corresponding to the original features and provides a foundation for estimating the contribution of each feature during attribution computation. Srinivas *et al.* [15] provided a theoretical discussion on the properties of baselines and demonstrates that reliable baselines must belong to the same open connected space (OCS) as the input (with further example in supplementary materials). This aspect has historically been overlooked because previous attribution research primarily focused on classification models trained on ImageNet, where objects are typically singular and centrally located in the image. This allows common feature expectations to construct a proper baseline, naturally ensuring that both the original features and the corresponding baseline belong to the same OCS, thereby maintaining high attribution quality. However, in autonomous driving scenarios, objects are scattered, leading to a feature map with strongly random spatial distributions. This randomness makes it challenging to establish a single reasonable baseline within the OCS. These observations suggest that generating a baseline requires identifying the OCS based on hidden layer information and

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, 25K03125). *(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)

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes detailed descriptions of the proposed algorithms and additional experimental results. This material is 10.1 MB in size.

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

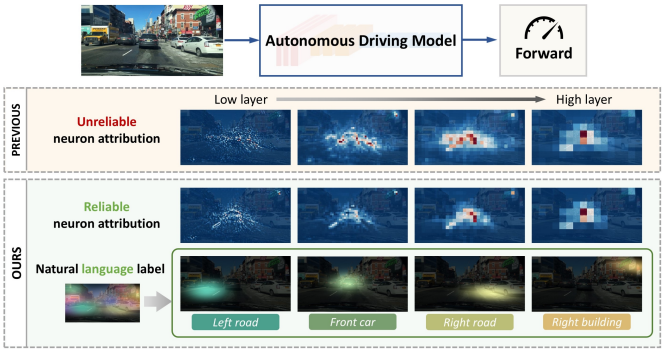


Fig. 1. Conceptual illustration of unreliable and reliable explanations. Previous methods produce inconsistent attribution results across different layers, frequently highlighting irrelevant regions. We aim for explanations that demonstrate consistency across layers and highlight the neurons influencing the decision-making process. Moreover, to enhance interpretability, incorporating natural language descriptions would be ideal.

applying specific constraints relevant to the driving scene.

Unreliable path. Attribution paths describe the trajectory from a baseline to an input, commonly defined as a straight line, by feature maps processed with Gaussian kernels, or through gradient-based paths. However, we find that these paths cannot ensure semantic coherence in the attribution task for autonomous driving models, leading to interpretations that highlight irrelevant regions and reduce attribution credibility. For example, when a straight-line path is used, the trajectory exhibits discontinuities within the same semantic region from the baseline to the input, and most sampling points along the path lack semantic discriminability, resulting in attribution noise over decision-irrelevant regions. This problem is less pronounced in classification models because the distribution of important semantic features is concentrated, and attribution results are less affected by paths. In contrast, the autonomous driving scenario presents a significant challenge due to the scale and random spatial distribution of semantic objects, making it difficult to achieve reliable attribution results with classical paths.

B. Solutions and Contributions

In response to the observations and discussions outlined above, we propose solutions to enhance the reliability of the AS baseline and path according to the properties of autonomous driving scenarios, ultimately leading to a more reliable attribution model.

Baseline solution. The key to solving baseline unreliability lies in designing OCS that can constrain the baseline based on the distribution of driving scenario features. Given the notable differences in feature distribution across different layers of an autonomous driving model, we introduce layer-specific upper and lower bound constraints to the baseline OCS. These bounds are informed by the expected extreme value distribution within the dataset. To further enhance baseline optimization, we propose two other constraints: gradient direction consistency and feature proximity. Our idea stems from the observation that a well-trained autonomous driving model should position an input near a local optimum on the loss surface. The baseline, representing the absence of information, should ideally align with the same gradient direction as the input but exhibit a larger magnitude (*e.g.*, a larger loss or gradient value). Furthermore, adhering to the fundamental definition of AS values, we introduce a feature proximity constraint, ensuring that the generated baseline remains close to the original input.

Path solution. The key to resolving the issue of unreliable integration paths lies in ensuring the semantic coherence of information along those paths. To achieve this, we use gradient-weighted feature maps to dynamically adapt the integration path. This strategy allows our attribution computation to effectively leverage semantically important regions within the feature maps. This is because gradient-weighted heatmaps exhibit a centro-diffusive distribution, with center points typically corresponding to semantic centers of objects or features. By incorporating gradient-weighted feature maps into the integration path definition, we ensure that the path remains confined within a consistent semantic region, which significantly reduces attribution noise arising from irrelevant regions.

To sum up, our main contributions are as follows:

- We identify two main problems when applying the Aumann-Shapley method to neuron attribution computation in common autonomous driving models: unreliable baseline selection and unreliable integration paths.
- We integrate the unique properties of autonomous driving data to define open connected space and gradient-weighted feature maps, proposing specific solutions and technical strategies for both reliable baseline generation and path creation. This approach enables robust and reliable computation of neuron attributions.

Our experiments incorporate a semantic self-supervised learning model to further enhance attribution explanations with natural language labels. An example is shown in Fig. 1, our attributions not only exhibit less noise and stronger consistency but also incorporate textual labels to provide a clear understanding of the semantic meaning behind the attribution results. We conduct quantitative and qualitative evaluations on a commonly used end-to-end autonomous driving model [17], [18] for explanation tasks. The results demonstrate that our method outperforms several state-of-the-art approaches. Our experiments extend beyond vision-only models to include a multimodal autonomous driving model (TransFuser [19]), incorporating diverse input modalities. The results demonstrate the general applicability and universality of our proposal.

II. Related Work

The explainability in autonomous driving has gained significant attention within the research community [20]–[22]. Recent literature broadly classifies attribution-based interpretation methods into three primary categories: perturbation-based approaches [23], [24], backpropagation-based techniques [25]–[30], and hybrid methods that combine elements of both [31], [32]. Our work focuses on enhancing the Aumann-Shapley method, a prominent technique within the backpropagation category, thereby contributing to the advancement of explainable autonomous driving.

Perturbation-based methods operate on the principle of analyzing how changes in the input features impact the model's output to determine feature attributions. Seminal work in this area includes Local Interpretable Model-agnostic Explanations (LIME) [33] and Deep Learning Important FeaTures (DeepLIFT) [24]. LIME generated perturbed inputs to train a simpler, interpretable linear model that approximates the original model's behavior locally. DeepLIFT quantified neuron activations relative to a reference state, effectively tracing the flow of influence between neurons. Recent research has focused on developing more sophisticated perturbation-based techniques capable of handling the increased complexity of DNNs. Sacha *et al.* [23] utilized semantic segmentation to adaptively identify key input features. Despite their flexibility, they often suffer from inefficiency due to the necessity of multiple iterations or operations for each input. Niu *et al.* [34] proposed a reinforcement learning-based framework, where the agent iteratively applies perturbation operations, such as masking or deleting tokens from the input sequence, to identify the tokens that are crucial to the decision-making process of large language models. Some recent studies indirectly leveraged the concept of ablation to quantify feature interactions or neuron contributions [35]–[37], but these methods often target specific model architectures, limiting their generalizability. Perturbation-based methods have made significant strides, but they generally suffer from high computational costs and are not well-suited for neuron-level attribution computation. This limitation is the primary motivation behind our introduction of a more efficient backpropagation-based method for neuron attributions in autonomous driving models.

Backpropagation-based methods have emerged as a dominant approach for explaining autonomous driving models, capitalizing on the accessibility of gradients with respect to input features. Saliency [38] directly uses gradients to highlight important features. Gradient Shapley (GradShap) [39] has been successfully applied to tasks like traffic object detection [27] and lane change prediction [25]. Zhang *et al.* [40] proposed concentration principle and designed SAMP for path attribution method, which searches the near-optimal path from a pre-defined path set. Chen *et al.* [28] proposed PropShapley, a technique that constructs a Shapley-value propagation model to facilitate efficient attribution computation. Wang *et al.* [41] proposed a simple Hierarchical Attribution Fusion (HAF) technique to enhance the smoothness of the optimized attribution heatmap. Liao *et al.* [42] proposed to dynamically visualize the attention flow by unlocking and accumulating spatial feature information produced by the class token across different Transformer layers. Furthermore, numerous DNN attribution methods [14], [29], [43]–[46], initially developed for diverse fields such as biology, economics, and psychology, could be adapted to enhance autonomous driving models with slight modifications.

Class activation mapping (CAM) [47] and GradCAM [48] represented another important gradient-based research direction, enabling attribution to a specific hidden layer. Recently, various CAM-based attribution methods have been developed to address different application scenarios. Techniques like BI-CAM [49], EigenCAM [50], HAG-XAI [1], and ODAM [51] have demonstrated success in handling diverse DNN architectures and achieving promising attribution results across multiple layers. CAM-based methods are generally characterized by their strong intuitiveness and low computational cost. However, their main limitation lies in the difficulty of achieving fine-grained, neuron-level attribution across all layers of the network, making it challenging to analyze the overall behavior of the neural network.

Hybrid attribution methods harness the strengths of both backpropagation and attribution baselines. Notable examples include layer-wise relevance propagation (LRP) optimization [31] and attribution aggregation techniques [32]. Ding *et al.* [52] leveraged the self-attention mechanism of Transformers to guide a multi-granularity random walk process, combining guidance based on the internal structure of the model with perturbation-based ideas. Choi *et al.*proposed ICEv2 [53], which attributes classification decisions to specific foreground patches by performing patch-level classification on Transformer output embeddings and learning via adversarial normalization. However, a critical limitation persists across many existing methods: while they effectively establish input-output causality, their approaches often lack generalizability. This means they require significant modifications depending on the specific scenario and model, limiting their applicability across different domains.

In addition to the attribution methods mentioned above, several studies have enhanced model interpretability or evaluated specific capabilities from different perspectives. Zhuang *et al.* [54] focused on generating natural language explanations for sarcasm phenomena in multimodal multi-party conversations. The core of this work lies in constructing a cross-modal collaborative guidance network rather than directly attributing predictions to input features. Xue *et al.*introduced ProVCIN [55], which integrates neural-symbolic reasoning with a variational causal inference network to generate coherent explanations for the reasoning process in explainable visual question answering, emphasizing reasoning chains rather than individual feature contributions. Du *et al.*proposed ST-Tree [56], a combination of Swin Transformer and neural tree models, which provides transparency for multimodal time-series classification through hierarchical decision paths in the tree structure. Huang *et al.* [57] introduced an innovative evaluation framework that utilizes the Japanese "Daigiri" game and causality-aware methods to assess the creative capabilities of multimodal large language models. These approaches enrich our understanding of complex models from multiple dimensions, including generating explanations, tracing reasoning processes, structural interpretability, and capability evaluation.

III. Aumann-Shapley Values in DNNs

To improve the explainability of DNNs traditionally viewed as black boxes, the adoption of Aumann-Shapley values rooted in game-theoretic principles has proven effective. Originally, Aumann-Shapley values are designed to assess the contributions of players in a game to a specific outcome. When adapted to neural networks, these values provide clear insights into how individual neurons influence the final output.

To ensure clarity, we begin by defining main terms. Let f represent the autonomous driving network, $x^{[0]}$ denote the input data, x represent the middle-layer feature maps under analysis, and $\bar{x}$ signify a baseline representing the absence of features. Typically, the input data $x^{[0]}$ is processed by the network f, resulting in the output $f(x^{[0]})$. The feature maps obtained at the middle layer l is denoted by $x = f^{[l]}(x^{[0]})$.

For a clearer representation of information embedded within hidden layers, we introduce a new notation $f(x)$. Here, $f(x)$ notation represents the network decision when the network is truncated at layer l, taking as input the neuron features x at the layer l. This definition simplifies the subsequent representation of attribution computation. The Aumann-Shapley approach focuses on evaluating the marginal contribution of each neuron, denoted as x_i, to the model's final decision d:

$$\phi_i = \int_{t=0}^{1} \left(f^d\left((1-t)\bar{x} + tx + \Delta x_i\right) - f^d\left((1-t)\bar{x} + tx\right)\right) dt, \tag{1}$$

where f^d stands for the d-th decision of the network. ϕ_i denotes the attribution of the i-th neuron, and $\Delta x_i = x_i - \bar{x}_i$. The term $(1-t)\bar{x}+tx+\Delta x_i$ represents the input transitioning from the baseline $\bar{x}$ to the target x, with adding influence from the feature x_i. As t progresses from 0 to 1, this formula traces a diagonal path that integrates x_i into the mix, highlighting its contribution to the output. Next, the first term of the integrand undergoes a Taylor series expansion:

$$\begin{aligned} f^d\left((1-t)\bar{x} + tx + \Delta x_i\right) &= f^d\left((1-t)\bar{x} + tx\right) \\ &+ \Delta x_i \frac{\partial f^d\left((1-t)\bar{x} + tx\right)}{\partial x_i} + O\left[(\Delta x_i)^2\right], \end{aligned} \tag{2}$$

where the remainder term $O\left[(\Delta x_i)^2\right]$ accounts for higher-order effects of the deviation. As $\bar{x}_i$ approaches x_i, this higher-order terms can be considered negligible. Thus, Eq. (1) becomes:

$$\phi_i = \Delta x_i \int_{t=0}^{1} \frac{\partial f^d\left((1-t)\bar{x} + tx\right)}{\partial x_i} dt. \tag{3}$$

The obtained Aumann-Shapley value ϕ_i is the gradient integral along the path between the baseline and the target input, quantifying the cumulative effect of the i-th neuron on the output prediction. The attribution results ϕ satisfy several key axioms of attribution explanation, ensuring that the obtained results are theoretically sound and meaningful (see supplementary materials for details).

IV. Reliable Aumann-Shapley Attribution

Despite the solid axiomatic foundation, the Aumann-Shapley attribution method often produces counterintuitive results when applied to autonomous driving scenarios. This raises a crucial question: How can attribution methods remain computationally aligned with foundational game theory principles while also provide intuitive and reliable feature attributions for autonomous driving models? Eq. (3) highlights that the selection of the baseline and the path from a feature-absent baseline state to a feature-present state are essential in the computational model. We will next focus on these two key aspects, analyzing and refining them within the context of autonomous driving applications. An overview of our proposal is shown in Fig. 2.

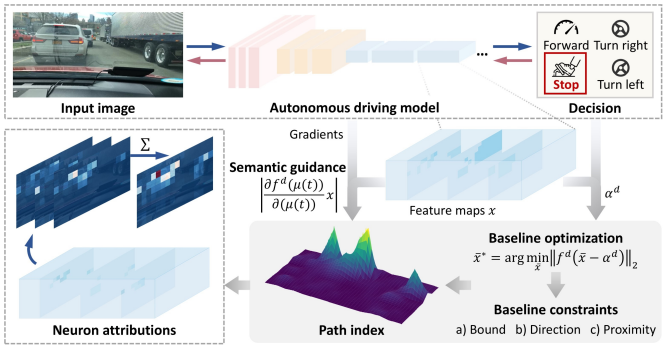


Fig. 2. A brief overview of our neuron attribution method. Blue arrows represent forward propagation, while red arrows indicate backward propagation. For this given input, the autonomous driving model's decision is to "stop." By constructing the baseline and integration path, we can obtain reliable attribution explanations.

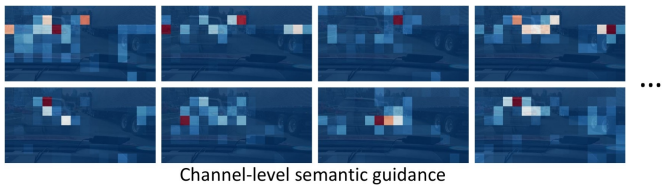


Fig. 3. Heatmaps of the top eight gradient-weighted feature maps. These results indicate that the weighted feature maps can effectively represent the traffic scene information related to the decision-making process.

A. Baseline Generation

A baseline's fundamental function is to represent the absence of decision-relevant features, serving as a reference for calculating attribution scores. Previously, baselines that reduce the decision score to zero are considered sufficient to signify missing information. However, this assumption does not hold true for autonomous driving models, where the four driving decisions ("forward," "stop," "turn left," and "turn right") are closely intertwined. Forcing one decision score to zero without considering others in such a system often introduces a significant number of adversarial features. Such a baseline not only fails to represent the absence of driving decision-relevant features but also directly misleads the attribution model.

To address this, we introduce the concept of equilibrium to generate a baseline. This definition ensures that the change in the original decision score introduced by a baseline should just lead to a corresponding shift in decision-making. In particular, instead of aiming for a zero score, we set a target score α^d that is just sufficient to induce a change in the original driving decision d. The basic objective function for baseline optimization can be defined as:

$$\bar{x}^* = \arg\min_{\bar{x}} \left\| f^d(\bar{x}) - \alpha^d \right\|_2, \tag{4}$$

$$\alpha^d = \mathbb{E}_{x \in \mathcal{X}^{\neg d}} \left[f^d(x) \right], \tag{5}$$

where $\|\cdot\|_2$ is the L2 norm. α^d represents the target score under the equilibrium state for the decision d. $\mathcal{X}^{\neg d}$ signifies the distribution of data samples excluding those where the decision is d. The overall optimization procedure relies on a simple yet effective stochastic gradient descent (SGD) approach. Having established the basic optimization objective, we next delve into the specific constraints required to further ensure baseline reliability in representing the absence of decision-relevant information.

During baseline optimization, the uneven distribution of features in driving scenarios may cause deviations in baseline information from the expected range and direction. To address this issue, we introduce open connected space (OCS) and restrict that both neuron features and the baseline belong to the same OCS. The theoretical foundations for this approach have been discussed in [14], [15], and we further clarify this concept within the specific context of autonomous driving. Specifically, we establish OCS for different channels within each layer l:

$$\delta^{[l]} = \left[\mathbb{E}_{x^{[0]}\sim\mathcal{X}}\left[\min\left(f^{[l]}(x^{[0]})\right)\right], \mathbb{E}_{x^{[0]}\sim\mathcal{X}}\left[\max\left(f^{[l]}(x^{[0]})\right)\right]\right], \tag{6}$$

where $\delta^{[l]}$ stands for the baseline value range constraint. $x^{[0]}$ represents an initial input sampled from the distribution $\mathcal{X}$, and $f^{[l]}$ denotes the output after the layer l. The $\min$ and $\max$ operations are applied spatially, across the feature map's height and width dimensions, while independently preserving each channel. If the baseline is outside the corresponding OCS, we update it by: $\bar{x}_i \leftarrow \bar{x}_i + \eta^{[l]}\mathrm{sgn}(x_i - \bar{x}_i)$, where $\eta^{[l]}$ means the step size and is calculated based on the range of OCS:

$$\eta^{[l]} = \frac{\left(\mathbb{E}_{x^{[0]}\sim\mathcal{X}}\left[\max\left(f^{[l]}(x^{[0]})\right)\right] - \mathbb{E}_{x^{[0]}\sim\mathcal{X}}\left[\min\left(f^{[l]}(x^{[0]})\right)\right]\right)}{m}, \tag{7}$$

where m is a predefined parameter that adjust the step size according the range of the OCS.

In addition to imposing range restrictions, we also apply a direction constraint to the baseline optimization in OCS, specifically ensuring that its gradient directions align with those of the neurons. This alignment is crucial for ensuring that the model's response to features is both coordinated and effective. If the gradient directions of the feature maps and the baseline are opposed, the baseline cannot represent the absence of features and could lead to discrepancies in the model responses. Therefore, during the optimization process, we evaluate the baseline direction by:

$$\mathrm{sgn}\left(\frac{\partial f^d}{\partial x_i}\right)\mathrm{sgn}\left(\frac{\partial f^d}{\partial \bar{x}_i}\right) < 0. \tag{8}$$

If this condition is satisfied, we proceed to update the baseline: $\bar{x}_i \leftarrow \bar{x}_i + \eta^{[l]}\mathrm{sgn}\left(\partial f^d/\partial x_i\right)$, making the gradient direction of the baseline coincide with that of the neuron.

Finally, according to the original definition of the Aumann-Shapley value, the baseline must remain close to the original neuron. To enforce this, we determine the threshold $\tau^{[l]}$ by analyzing the probability distribution of neuron values obtained from the autonomous driving dataset. The tenth percentile of this distribution serves as the threshold. During optimization, if the absolute difference between the baseline and the neuron exceeds this threshold, we skip that index from further updates to prevent significant deviations (see supplementary material for the algorithm details). With these designed constraints implemented, we establish a reliable baseline that faithfully represents the absence of specific features within the decision-making process.

B. Path Selection

After establishing the baseline, the next critical step involves selecting the integration path. Ideally, the path should flow from the absence to the presence of features, ensuring the preservation of the desired characteristics integral to the path attribution framework. In the original Aumann-Shapley values, the integration path $\mu'(t) = (1-t)\bar{x} + tx$ follow a straight line in the feature space. While this path can be effective for games that can be modeled with linear functions, allowing for near-perfect sampling between the baseline and the input, it is not optimal for complex scenarios.

For autonomous driving task, conventional choices like linear paths and gradient paths prove unsuitable. These model-agnostic paths fail to maintain semantic coherence; they cause abrupt transitions from the baseline to the input within the same semantic region, therefore breaking the surface continuity of the computational model. This breakage leads to unreliable and noisy attributions at various locations within the autonomous-driving scenarios.

To preserve the integrity of semantically coherent regions during integration, we introduce a saliency-guided approach for defining the integration path. Our method utilizes gradient-weighted feature maps, which approximate the contribution of different features and offer critical semantic insights. Gradient-weighted saliency highlights semantic regions corresponding to important objects across different channels, expanding outward from the object's center to include its surrounding context, as shown by the 3D mesh in Fig. 2. By employing gradient-weighted feature maps as a path flow index, our approach enables dynamic adaptation of the integration path according to the input data. Eight examples of gradient-weighted feature maps are shown in Fig. 3, where it can be observed that the gradient-weighted saliency already provides a rough indication of the traffic scene information driving the decision-making process. As a result, the integration paths are constrained within semantically consistent regions, thereby ensuring coherence. Formally, this constraint can be expressed as:

$$\mathcal{L}_{\text{cohr}} = \left\| \int_{t=0}^{1} \left| \frac{\partial f^d\left(\mu(t)\right)}{\partial\left(\mu(t)\right)} x \right| dt \right\|_2, \tag{9}$$

where $\mu(t)$ is the path from the baseline to the input. By minimizing $\mathcal{L}_{\text{cohr}}$, we increase the semantic concentration of the generated attributions and simultaneously reduce irrelevant noise, thereby enhancing the reliability and clarity of the interpretations. To ensure that the solution path remains bounded

and to find the optimal trajectory, we introduce an additional distance constraint. This constraint is designed to keep the path distance close to that of a straight-line path:

$$\mathcal{L}_{\text{dist}} = \int_{t=0}^{1} \|\mu'(t) - \mu(t)\|_2 \, dt, \tag{10}$$

where $\mu'(t)$ denotes a straight-line path from the baseline to the input. This constraint has the added benefit of enhancing stability and computational efficiency during the optimization process.

To achieve the above objectives, we introduce specific strategies for efficient implementation. Initially, we split the straight-line path from the baseline to the input into K segments. For each of these segments, we define a set $\mathbb{S}^{(k)}$, which includes indices of neurons whose values have not yet reached the input values:

$$\mathbb{S}^{(k)} = \left\{ i | \hat{x}_i^{(k)} \neq x_i \right\}, \tag{11}$$

where $\hat{x}_i^{(k)}$ denotes the current value of neuron i within segment k. With this set established, instead of simultaneously advancing all neurons towards their input values, we selectively focus on a subset $\mathbb{U}^{(k)}$. This subset is chosen based on the designed criteria $h_i^{(k)}$, where we filter neurons by the semantic coherence:

$$\mathbb{U}^{(k)} = \left\{ i \mid h_i^{(k)} \leq \text{quantile}(\{h_i^{(k)} \mid i \in \mathbb{S}^{(k)}\}, \alpha) \right\}, \tag{12}$$

$$h_i^{(k)} = \left| \frac{\partial f^d(x)}{\partial x_i} x \right|, \tag{13}$$

where α specifies the quantile threshold used for neuron selection. For instance, if $\alpha = 0.1$, we select the neurons corresponding to the lowest 10% of $h_i^{(k)}$ values from the set $\mathbb{S}^{(k)}$ to form the subset $\mathbb{U}^{(k)}$. Subsequently, we progressively move the neurons in the subset across each segment. Guided by the gradient-weighted feature maps, the entire integration path ensures a semantically coherent transition from the baseline to the original features. This effectively mitigates attribution noise in irrelevant regions, ultimately enhancing the reliability of the calculated attributions. Note that, in multimodal networks, each modality typically corresponds to a distinct branch, and the dimensions of neuron features are not identical. Consequently, paths should be computed separately for each modality during generation. Furthermore, the numerical distributions across different modalities often vary significantly; handling them independently also contributes to maintaining numerical stability throughout the generation process. Further implementation details and parameter settings can be found in supplementary materials.

V. Experimental Results

A. Comparative analysis of attribution methods

The proposed methodology is designed to ensure the reliability of attribution computation. In this section, we leverage qualitative comparisons to visually show the advantages of our proposal. We also evaluate the reliability and reasonableness of our attribution results using multiple evaluation metrics.

Our experiments encompass a comparison with several state-of-the-art attribution methods: Saliency [38], GradShap [27], IDGI [29], Guided GradCAM (GGCAM) [58], PropShap [28], SAMP [40], and LRP-*h* [31]. Saliency serves as a representative and important baseline, embodying a classic attribution approach. GradShap, IDGI, PropShap, and SAMP are all rooted in Shapley values, each representing distinct Shapley value estimation models. GGCAM, an improved variant of GradCAM, is also included in our evaluation as it represents a significant direction in attribution research. These methods can be applied to autonomous driving models with minimal modifications. In contrast, LRP presents a unique case. Currently, numerous propagation rules for LRP have been developed. To achieve reasonable results within our context, we find it necessary to employ a combination of LRP techniques tailored to specific model components. The hybrid LRP approach in our experiments, denoted as LRP-*h*, involves using LRP-0 at the decision layer, LRP-ϵ in higher layers, and LRP-γ in lower layers. For regularization and non-linear layers, our implementation primarily draws upon the strategies outlined in [31], [59], [60].

We conduct our comparative experiments on the output layers of the four main blocks within DenseNet (denoted as Block 1-4 from low layer to high layer). Each block incorporates a downsampling operation, representing the primary feature abstraction process in DenseNet. These layers can be considered critical for characterizing the feature representation within the network.

Fig. 4 presents a qualitative comparison of attribution heatmaps, which are generated by aggregating neuron attributions across the channel dimension. The top-right corner of each original image displays the model's decision, encompassing "forward," "stop," "turn left," and "turn right." Each sample is accompanied by four rows of heatmaps, representing attributions from lower to higher blocks (from top to bottom). We can find our method effectively suppresses various types of attribution noise while exhibiting superior cross-layer consistency. In contrast, other methods, even those demonstrating comparable quantitative performance, often fall short in this regard. They produce attributions where salient features highlighted in lower layers vanish in higher layers, introducing inconsistencies that can undermine the reliability of attribution results.

For quantitative evaluation, we select five attribution metrics: Sensitivity-n (Sen-n) [26], [61], [62], Accuracy Information Curve (AIC), Softmax Information Curve (SIC) [32], Least-Relevant-First (LeRF), and Most-Relevant-First (MoRF) [61]. Sensitivity-n serves as a primary indicator of attribution effectiveness and is crucial for ensuring theoretical soundness. AIC, SIC, LeRF, and MoRF share a similar design philosophy, aiming to quantify how well the attributions identify truly important features. AIC and SIC define this evaluation from an information entropy perspective, while LeRF and MoRF directly leverage decision scores for a more direct assessment. Collectively, these metrics provide a multifaceted view of whether the attributions accurately identify critical neuron features, thereby assessing the reliability and reasonableness of the numerical attribution values.

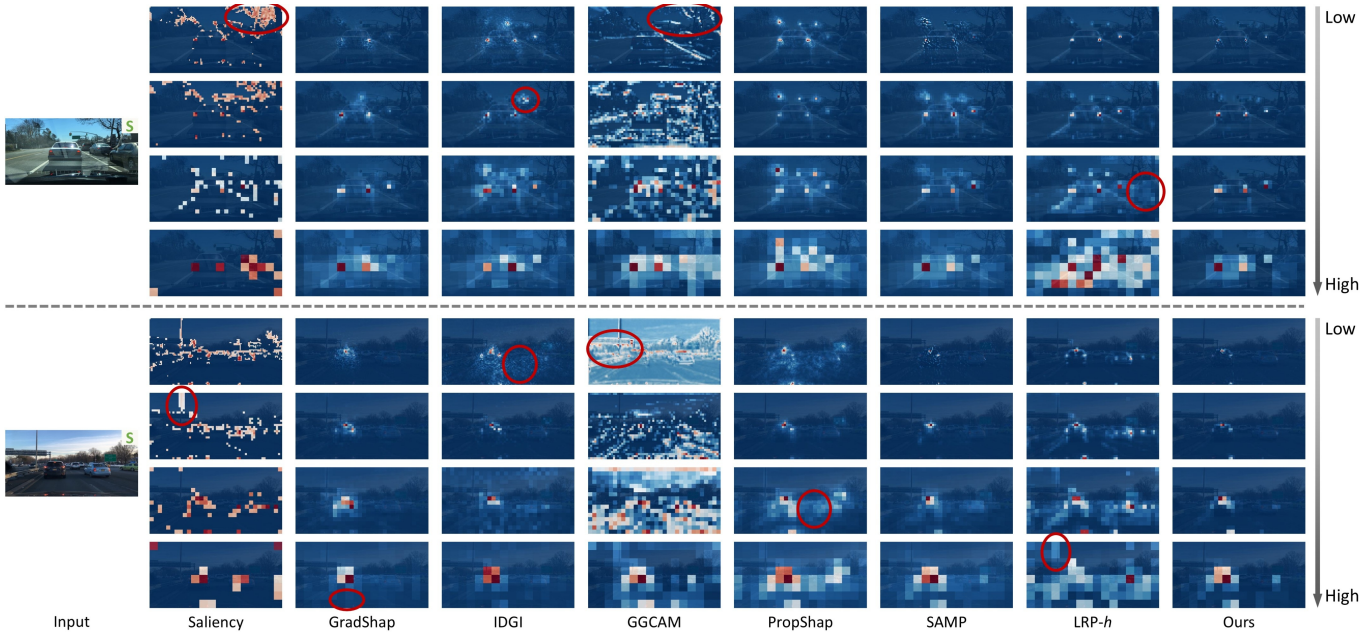


Fig. 4. Qualitative attribution comparison between ours and other methods across different layers.

TABLE I
QUANTITATIVE COMPARISON BETWEEN OURS AND OTHER METHODS ACROSS DIFFERENT LAYERS.

	Block 1					Block 2				
	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
Saliency	0.622	0.572	0.597	0.763	0.206	0.614	0.564	0.587	0.745	0.211
GradShap	0.748	0.624	0.645	0.827	0.131	0.776	0.611	0.642	0.789	0.164
IDGI	0.716	0.596	0.606	0.804	0.172	0.762	0.614	0.639	0.781	0.159
GGCAM	0.603	0.567	0.582	0.742	0.209	0.582	0.552	0.572	0.738	0.232
PropShap	0.752	0.636	0.654	0.837	0.133	0.788	0.629	0.654	0.812	0.157
SAMP	0.769	0.662	0.682	0.849	0.102	0.811	0.659	0.683	0.804	0.132
LRP-*h*	0.738	0.611	0.632	0.811	0.167	0.761	0.599	0.617	0.774	0.171
Ours	**0.792**	**0.702**	**0.724**	**0.875**	**0.087**	**0.833**	**0.696**	**0.717**	**0.862**	**0.103**
	Block 3					Block 4				
	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
Saliency	0.627	0.567	0.586	0.726	0.239	0.821	0.562	0.583	0.751	0.194
GradShap	0.744	0.621	0.635	0.757	0.159	0.854	0.603	0.622	0.788	0.179
IDGI	0.736	0.606	0.619	0.758	0.178	0.856	0.589	0.605	0.782	0.153
GGCAM	0.597	0.548	0.569	0.702	0.255	0.842	0.566	0.589	0.779	0.177
PropShap	0.751	0.629	0.649	0.778	0.162	0.871	0.619	0.641	0.804	0.132
SAMP	0.775	0.656	0.674	0.792	0.131	0.889	0.641	0.671	0.812	0.129
LRP-*h*	0.715	0.591	0.627	0.749	0.194	0.719	0.544	0.567	0.713	0.215
Ours	**0.817**	**0.691**	**0.714**	**0.847**	**0.091**	**0.914**	**0.673**	**0.691**	**0.833**	**0.112**

The quantitative result in Table I reveals several noteworthy patterns. Sensitivity-n, which measures how well the attribution values satisfy the "efficiency" axiom, generally exhibits better performance in hidden layers closer to the output. This trend is often attributed to the reduced impact of non-linear perturbations in layers closer to the output. AIC and SIC assess the ability to progressively recover neuron features based on the assigned attribution values, indirectly reflecting the accuracy of attributions in capturing the contributions of critical features. LeRF and MoRF provide a more direct evaluation of the rationality of the numerical attribution distribution.

Interestingly, we observe that in Block 4, the quantitative results do not align well with the patterns seen in shallower layers. For example, LRP exhibits a significant decline in performance on the MoRF metric. LRP results in higher layers can be obtained through only a few propagation operations; however, due to the design of its propagation rules, it tends to focus on the overall distribution of positive contributions. This leads to a lack of emphasis on the most critical features. When a model's decision relies on the absence of certain features as critical information, LRP may fail to identify these as most relevant features, thereby not prioritizing their removal under the MoRF paradigm. This specific example also underscores a broader point: current attribution metrics, by primarily evaluating outcomes after feature removal, may not always fully capture the features' actual, and often nuanced, contributions to the model's original decision.

In stark contrast to some of these observed limitations, our

TABLE II
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS.

	Block 1					Block 2				
	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
Straight line path	0.765	0.654	0.671	0.843	0.109	0.804	0.661	0.674	0.817	0.127
Random seg. path	0.763	0.657	0.679	0.836	0.111	0.812	0.654	0.677	0.819	0.126
Curve interpolat. path	0.739	0.662	0.685	0.848	0.107	0.809	0.668	0.679	0.828	0.118
$K = 1$	0.749	0.623	0.641	0.809	0.132	0.788	0.624	0.637	0.791	0.154
$K = 5$	0.783	0.692	0.711	0.859	0.095	0.819	0.679	0.701	0.851	0.110
$K = 20$	0.788	**0.703**	0.721	0.871	0.090	0.826	0.693	0.714	0.859	0.106
$\alpha = 0.05$	0.791	0.699	0.720	0.869	0.089	0.831	0.693	0.716	0.858	0.105
$\alpha = 0.15$	0.788	0.698	0.719	0.872	0.089	0.829	0.691	0.713	0.855	**0.102**
$\alpha = 0.2$	0.786	0.688	0.707	0.866	0.092	0.825	0.683	0.704	0.851	0.109
Zero baseline	0.761	0.649	0.663	0.836	0.115	0.798	0.646	0.669	0.803	0.139
Uniform baseline	0.719	0.611	0.639	0.793	0.143	0.782	0.616	0.628	0.781	0.163
Gaussian baseline	0.732	0.625	0.642	0.811	0.136	0.791	0.629	0.642	0.792	0.149
Max dist. baseline	0.713	0.602	0.624	0.779	0.183	0.763	0.595	0.609	0.771	0.172
Ours	**0.792**	0.702	**0.724**	**0.875**	**0.087**	**0.833**	**0.696**	**0.717**	**0.862**	0.103
	Block 3					Block 4				
	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓	Sen-n ↑	AIC ↑	SIC ↑	LeRF ↑	MoRF ↓
Straight line path	0.791	0.664	0.681	0.803	0.126	0.904	0.657	0.678	0.817	0.125
Random seg. path	0.796	0.657	0.679	0.809	0.119	0.907	0.641	0.679	0.812	0.122
Curve interpolat. path	0.802	0.669	0.685	0.819	0.112	0.909	0.659	0.681	0.816	0.118
$K = 1$	0.763	0.631	0.667	0.786	0.158	0.885	0.632	0.647	0.788	0.148
$K = 5$	0.806	0.682	0.701	0.831	0.097	0.907	0.665	0.685	0.824	0.119
$K = 20$	0.815	0.688	0.711	0.843	0.094	0.912	0.669	0.689	0.829	0.116
$\alpha = 0.05$	0.816	**0.694**	0.713	0.845	**0.091**	0.911	0.672	0.689	0.831	0.113
$\alpha = 0.15$	0.814	0.689	**0.718**	0.843	0.093	**0.915**	0.671	0.688	0.832	0.115
$\alpha = 0.2$	0.813	0.687	0.705	0.836	0.095	0.909	0.668	0.685	0.825	0.120
Zero baseline	0.783	0.657	0.676	0.811	0.129	0.891	0.649	0.669	0.811	0.122
Uniform baseline	0.751	0.622	0.631	0.768	0.161	0.866	0.621	0.639	0.785	0.156
Gaussian baseline	0.766	0.625	0.649	0.781	0.154	0.872	0.634	0.651	0.784	0.146
Max dist. baseline	0.746	0.602	0.612	0.754	0.169	0.854	0.607	0.627	0.767	0.172
Ours	**0.817**	0.691	0.714	**0.847**	**0.091**	0.914	**0.673**	**0.691**	**0.833**	**0.112**

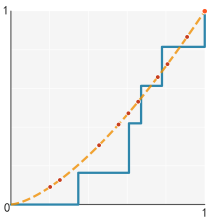


Fig. 5. A comparative illustration of attribution integration paths. The blue polyline represents the random segmentation path. The orange curve represents the curve interpolation path.

proposed method consistently achieves the best performance across all these diverse metrics. Particularly compelling is its performance on the MoRF metric, where our approach demonstrates a nearly 50% enhancement over the baseline Saliency method. This substantial gain strongly highlights our method's accuracy in identifying the most influential neuron features, as their removal correctly leads to significant alterations in the model's decision scores.

B. Ablation Studies

This section details ablation experiments conducted on our proposal, with a primary focus on two components: path selection and baseline. For path selection, three additional paths are evaluated. The straight-line path corresponds to the original definition in Aumann-Shapley values, connecting the baseline to the original features in a direct line. To validate the proposed method's performance, two distinct multi-path approaches are also designed. For the random segmentation path, a randomly generated segmented path approach is adopted. In this approach, the parameter t in Eq. (3) is randomly discretized into 10 segments, and 20 groups of paths are randomly generated to compute the average attribution. This path is illustrated by blue polylines in Fig. 5. For the curve interpolation path, we randomly sample 10 points on one random curve, generate 20 groups of random curves, and compute the average attribution results. The path illustration is shown by the orange curve in Fig. 5. Although multi-path approaches improve attribution results compared to the straight-line path, they still introduce some unknown noise in irrelevant regions, as shown in Fig. 6. The qualitative results also demonstrate that our method produces less noise compared to other alternatives.

Additionally, we design ablation experiments for two key parameters in our path selection: the segment number K and the threshold condition parameter α. First, we hold α constant at 0.1 and test K at values of 1, 5, 10 (our setting), and 20. Subsequently, keeping K at 10, we compare α values of 0.05, 0.1 (our setting), 0.15, and 0.2. The results in Table II indicate that K exhibits a clear marginal effect at a value of 10; further increases typically yield no additional improvements. A similar pattern occurs for α at 0.1, where both larger and smaller thresholds slightly diminish the quality of the attribution results. Increasing the number of segments or lowering the threshold, which represents finer discretization,

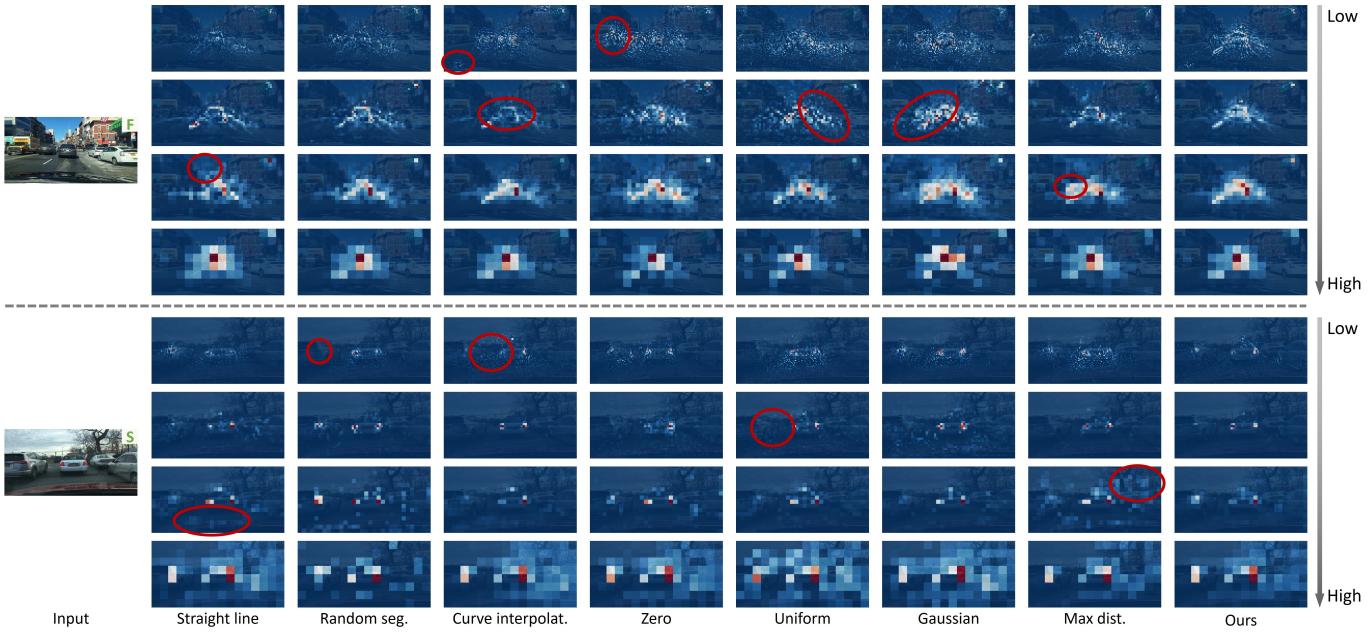


Fig. 6. Qualitative comparison results of ablation experiments.

Fig. 7. Attribution results on the original noise-free image.

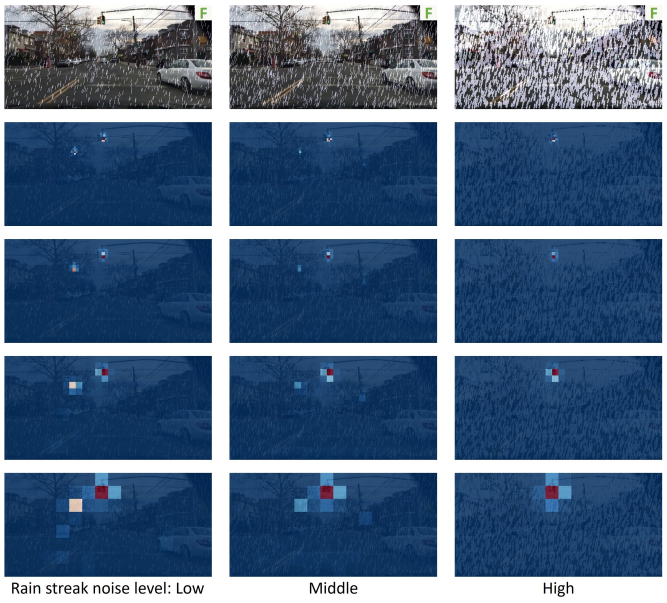


Fig. 8. Attribution results on images perturbed by rain streaks.

does not show significant improvements.

Finally, we incorporate four different baselines for comparison, drawing from literature [26], [63] on Shapley value baselines: zero, uniform noise, Gaussian noise, and max distance baseline. To account for the impact of random noise, we generated 20 baseline samples and calculated the average attribution across these baselines to obtain the final attribution result. Our findings indicate that random noise baselines (uniform and Gaussian) and the max distance baseline typically produce attribution results inferior to the conventional zero baseline. A possible reason is that, for neurons, a zero value might offer better generalizability when representing missing information. The max distance baseline can be considered a model-specific selection, merely modifying neuron values based on their numerical distribution to achieve maximum distance does not appear to effectively represent feature absence. As shown in Fig. 6, these three alternative baselines often highlight irrelevant regions. Furthermore, their quantitative results, presented in Table II, are obviously weaker than those of alternative methods evaluated.

C. Attribution Robustness under Data Perturbations

In this section, we evaluate the robustness of the attribution method under data noise perturbations. We apply various levels of rainy conditions, foggy conditions, and random noise to the input data, and subsequently generate corresponding attribution explanations. The attribution results for the original images, as well as those with each type of noise, are presented in Fig. 7 to Fig. 10.

We observe that the attribution results effectively reveal the primary causes underlying the model's decisions. However, as the noise intensity increases and information from the original image is progressively lost or obscured, the attribution results no longer consistently highlight the corresponding regions. For instance, under high-intensity rainy conditions where a traffic light is completely obscured, the attribution ceases to highlight

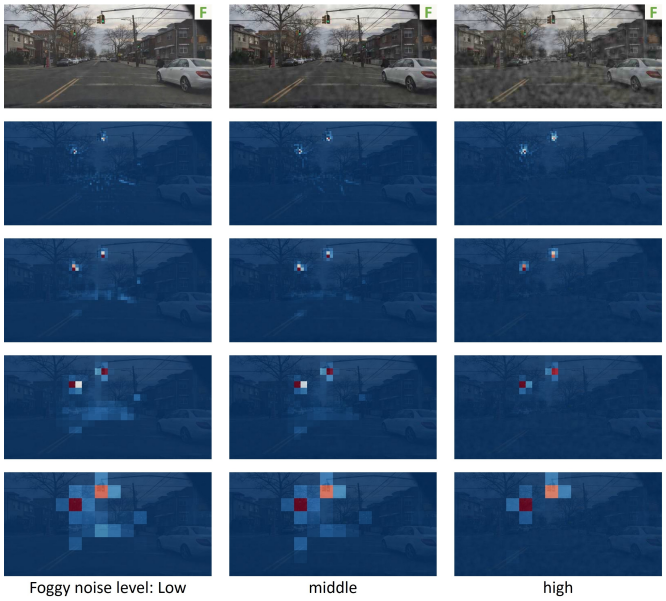


Fig. 9. Attribution results on foggy images.

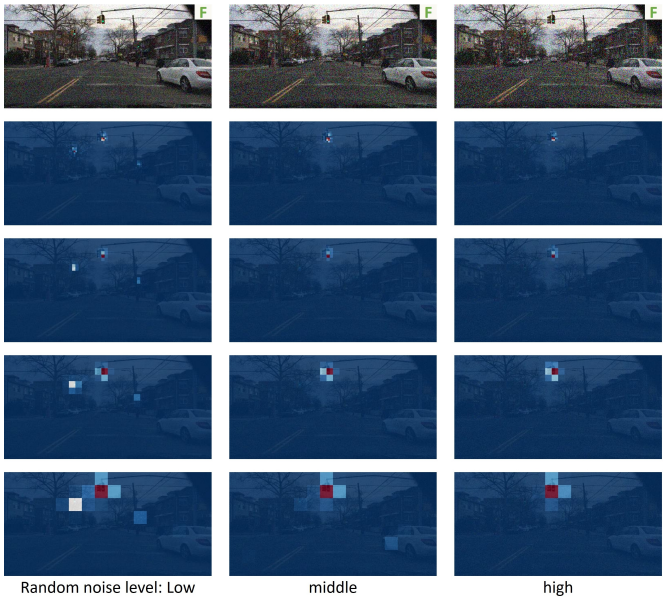


Fig. 10. Attribution results on images perturbed by random noise.

this region (Fig. 8). Additionally, we find that as random noise intensity increases, the attribution results also exhibit changes. For example, after applying medium and high random noise, two traffic lights are still roughly visible in the image, but the attribution results no longer highlight the second traffic light (Fig. 10).

These experiments demonstrate that our proposal exhibits a commendable level of robustness. Even when subjected to noise interference, our method can often still identify the key factors driving the model's decisions. However, when noise corrupts critical information within the original data, the network itself fails to detect the corresponding information. This leads to significant differences between the attribution results for the noisy images and those for the original images.

Fig. 11. Attribution explanations with nature language semantic labels. The semantic blobs are ordered according to their corresponding attribution scores.

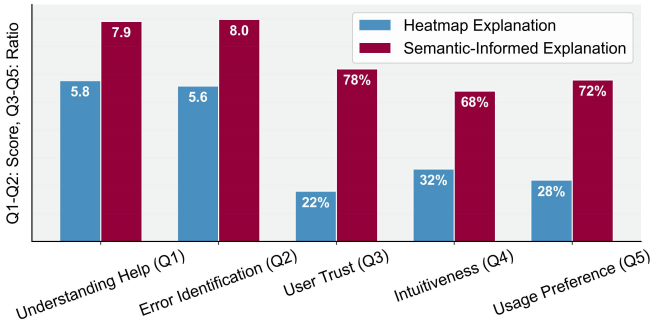


Fig. 12. User study results comparing semantic-informed attribution explanations with original numerical attribution explanations.

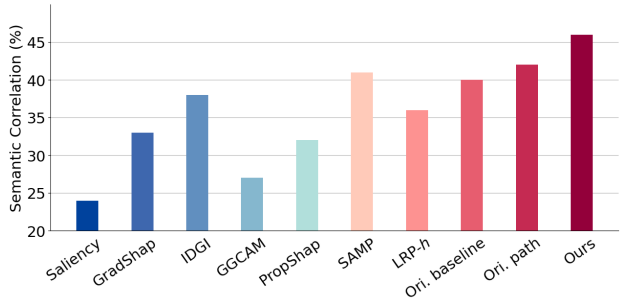


Fig. 13. Attribution semantic correlation ratio. Higher values signify that the attribution results are more focused within semantic segmentation regions critical to decision-making, as defined by the semantic labels of the BDD dataset.

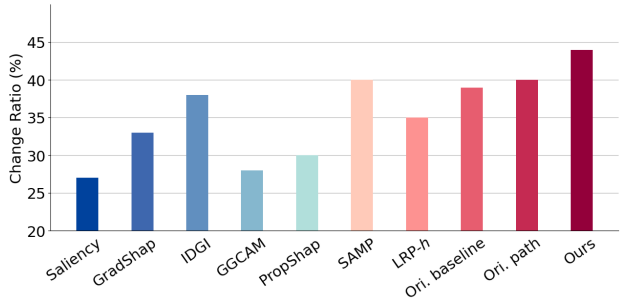


Fig. 14. Decision score change ratio. Higher values indicate the attribution results more accurately pinpoint the key semantics influencing the decision-making process.

D. Semantic-Informed Attribution Explanation

In this section, we leverage BlobGAN [64] to enhance the semantic expressiveness and interpretability of our attribution results. A common criticism of attribution explanations is their inherent lack of clarity; highlighted regions in a heatmap often lack clear semantic meaning. By incorporating natural language semantics through BlobGAN, we improve the interpretability of these explanations, providing a more readily understandable context for the highlighted regions.

Specifically, we train a BlobGAN model with BDD dataset using the unsupervised training settings discussed in [64]. This model utilizes an encoder to transform images from autonomous driving scenarios into semantic blobs, and includes a generator capable of reconstructing these blobs into images. Blobs generated through extensive unsupervised training exhibit correlations with specific objects within traffic scenes.

With these semantic-related blobs, we then use the semantic segmentation labels from the BDD dataset to determine the intersections between the pixel regions impacted by each blob and their corresponding semantic segmentation areas. This step enables us to assign natural language semantics to each blob. Additionally, we manually add directional labels such as front, left, and right, to each blob. In particular, when a blob is removed, its corresponding pixel region is affected. We calculate the overlap between the pixel regions influenced by the blobs and the semantic labelled regions. Blobs with the largest overlap across the dataset are identified as the semantic components corresponding to the semantic labels. Using semantic blobs, we can directly extract and interpret the semantic information embedded in the attribution values by calculating their intersection, which facilitates a deeper understanding of model decisions and behaviors. Fig. 11 shows the generated attribution explanations with text-level labels.

To investigate whether the inclusion of semantic information improves the interpretability of attribution explanations, we conduct a comparative experiment involving 46 participants from three universities. These participants have only basic knowledge of artificial intelligence and are unfamiliar with autonomous driving and attribution methods. A crossover design was used to mitigate potential learning and fatigue effects: half of the participants first evaluated the original attribution explanations followed by the semantic-informed explanations, while the other half evaluated them in the reverse order, with the presentation order of image samples randomized for all participants.

For each sample, participants were asked to review the model predictions and corresponding explanations, and subsequently provided subjective evaluations based on the following questions: (1) To what extent does the explanation help you understand the model's decision-making basis (Understanding Help)? (2) In cases of incorrect model decisions, to what extent does the explanation assist in identifying potential model errors (Error Identification)? (3) Which explanation method better instills your trust in the model's predictions (User Trust)? (4) Which explanation do you find more intuitive and easier to understand (Intuitiveness)? (5) Which explanation would you prefer to use (Usage Preference)? The experimental results are summarized in Fig. 12, wherein Q1 and Q2 show the scores, while Q3 to Q5 present the preference percentages for each explanation method. The result shows that semantic-informed explanations are consistently preferred across all criteria. For example, user trust and intuitiveness increased by more than double, and over 70% of participants chose the semantic explanations as their preferred interpretation method.

To further evaluate the semantic representational capacity of attribution methods quantitatively, we introduce two metrics. The first metric leverages semantic segmentation labels. We begin by summing the attribution results from all tested methods to obtain a comprehensive spatial attribution distribution. This distribution guides us in identifying the most influential semantic labels. Next, using the ground-truth segmentation labels, we determine the pixels associated with each semantic label within an image. By calculating the intersection between these pixels and the attribution heatmap, we can quantify the alignment between attributions and relevant semantics. Specifically, the ratio of the sum of attribution values within this intersection to the total attribution value serves as our evaluation metric. A higher ratio indicates stronger semantic correspondence. As shown in Fig. 13, our method demonstrates superior semantic correspondence compared to all other tested methods.

However, this evaluation metric relies on datasets with segmentation labels, which are often unavailable for autonomous driving datasets. Therefore, we further introduce a clustering-based evaluation approach that eliminates the dependency on segmentation labels. We first employ mean-shift clustering to group attribution values exceeding the 0.1% quantile. For each resulting cluster, we identify the nearest semantic blob. By removing these blobs, we effectively eliminate the corresponding semantic information from the input. We then measure the resulting change ratio in the decision score. A significant change suggests that the removed semantics are indeed influential, indicating a strong semantic correspondence for the attribution method. As illustrated in Fig. 14, our method outperforms other comparative methods under this clustering-based evaluation as well.

E. Feature Attribution Distribution Analysis

In this section, we explore the relationship between feature activations and their contributions to decision-making using joint distribution plots of feature values and corresponding attributions. We examine the common assumption that higher activation implies higher contribution when using ReLU activations. This assumption, often lacking robust evaluation, is directly addressed here through our joint distribution analysis. Our findings reveal a stronger correlation between activation and contribution in hidden layers closer to the output. This correlation weakens in earlier layers, where even low-activation neurons can exhibit significant contributions, likely amplified through the network forward propagation. This also suggests that the activation of higher-layer neurons, having undergone multiple transformations and thus becoming more discriminative, serves as a more direct indicator of their contribution. Analyzing the joint distribution facilitates the identification of these crucial features.

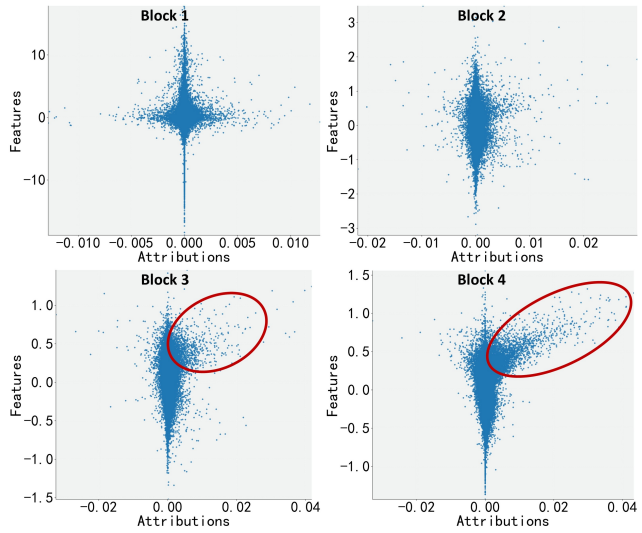


Fig. 15. Joint scatter plot of feature values and attributions among different blocks on the DenseNet-based driving model [17].

Fig. 16. Attribution results generated on the second feature level of TransFuser model. Given the model's multimodal input (image and LiDAR data), the attributions are correspondingly divided into two parts, reflecting the contributions of each modality.

Previous feature analysis, relying solely on activation levels to interpret model behavior, risk overlooking important contributing factors. By incorporating joint distribution plots of feature activations and attributions, we provide a more detailed understanding of feature importance, capturing contributions that might otherwise be missed when focusing solely on activation magnitude.

F. Evaluation on Multimodal Network

To validate the general applicability of our method, we implement attribution computation on a complex, multimodal autonomous driving model, TransFuser [19]. This model is representative of state-of-the-art architectures, incorporating visual and LiDAR Bird-Eye View (BEV) inputs, utilizing a Transformer architecture for feature fusion, and comprising eight different sub-modules, including a sequential processing module.

Existing attribution methods often struggle with the increasing complexity of multimodal autonomous driving models, which typically feature multiple sub-modules. Our proposed method enables the independent computation of different modality branches, thereby effectively avoiding numerical instability issues that can arise when processing neurons with substantially divergent value distributions. By distinctly calculating the baselines and integration paths for each modality, the accuracy of the attribution is reliably preserved. Furthermore, the diversity of output types in these models can distort attribution evaluation metrics, hindering effective comparison of different attribution results. This challenge limits the ability to conduct effective comparative attribution experiments and represent an important area for future research.

Our experiments use the pre-trained weights and dataset provided in the original paper [19]. Fig. 16 shows the attribution results calculated at the model's second feature level, with attribution dimensions of 88×20 for the image branch and 32×32 for the BEV branch. The results demonstrate TransFuser's clear feature extraction capabilities, with the image and LiDAR branches focusing on different aspects of the driving environment, showcasing the effectiveness of Transformer-based feature fusion. From the attribution results, we observe that the influence of different modalities on the decision-making process appears to be collaborative. However, current attribution explanations lack the capacity for interactive evaluation, and investigating the interactive effects across different modalities remains an open challenge.

VI. Conclusion

In this work, we introduce a novel attribution approach designed to provide reliable explanations for autonomous driving models. Motivated by the discovery of counterintuitive results when applying standard path integral method directly to driving scenarios, we pinpoint the primary sources of shortcomings within the attribution framework that compromise its reliability. To address these problems, we propose attribution baselines and paths tailored to the specific characteristics of the autonomous driving scenario. Our design adheres to original game-theoretic criteria and accounts for the decentralized nature of traffic object distribution, ensuring that the process from the baseline along the path to the original features accurately reflects the feature transition. Our experiments highlight the effectiveness of the proposed approach across multiple attribution metrics and visualization effects.

Despite the progress made, several open questions remain worthy of discussion. Firstly, the baseline generation process in our method, while effective in ensuring attribution reliability, also introduces obvious computational overhead. This is primarily due to the necessity of an initial inference pass to gather expectation information from the entire dataset, and the application of multiple constraints during the optimization process. In our experiments, generating a single robust baseline typically takes 26.5 seconds. When using a high-performance GPU and a larger batch size (*e.g.*, 128), the computational efficiency can be improved by approximately two orders of magnitude. However, this additional baseline generation step is not required for conventional methods that employ fixed or random baselines. This computational overhead not only presents a cost issue but also makes it more challenging to isolate and analyze the influence of the dataset's feature distribution on the final attribution results. Consequently, the trade-off between computational cost and attribution accuracy

becomes a key consideration, especially for real-time applications. Although our current approach prioritizes generating robust and interpretable attributions, future research could explore approximate baseline techniques or pre-computed baselines tailored to specific scenarios. These strategies hold the promise of significantly saving computational resources without excessively compromising attribution accuracy.

Beyond the challenges in computational efficiency, evaluating attribution effectiveness also presents critical open questions. Existing attribution evaluation metrics struggle to adequately assess the latest, increasingly complex multimodal autonomous driving models. Designing new evaluation metrics that consider the features used in attribution computation, as well as the model's input and output characteristics, is an important direction for future research. Furthermore, integrating attribution methods with evaluation metrics specific to autonomous driving, such as hazard detection rates or misjudgment rates, remains a significant challenge. Addressing this issue represents a valuable direction for advancing research in autonomous driving and could further enhance the practical value of attribution explanations.

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] L. Zhao, H. Zhou, X. Zhu, X. Song, H. Li, and W. Tao, "LIF-Seg: LiDAR and camera image fusion for 3D LiDAR semantic segmentation," *IEEE Trans. Multim.*, vol. 26, pp. 1158–1168, 2024.

[3] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Traffic scene-informed attribution of autonomous driving decisions," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 7, pp. 9175–9186, 2025.

[4] M. Li, H. Sun, Y. Huang, and H. Chen, "Svap: Shapley value guided attribution prior for neural network-based autonomous driving," *IEEE Trans. Veh. Technol.*, pp. 1–12, 2025.

[5] M. Li, Z. Cui, Y. Wang, Y. Huang, and H. Chen, "An explainable q-learning method for longitudinal control of autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 3, pp. 4214–4218, 2025.

[6] L. Wang, S. Sun, and J. Zhao, "VirPNet: A multimodal virtual point generation network for 3D object detection," *IEEE Trans. Multim.*, vol. 26, pp. 10 597–10 609, 2024.

[7] J. Zhan, L. Zhang, and J. Qiao, "Boundary consensus of networked hyperbolic systems of conservation laws," *IEEE Trans. Autom. Control*, pp. 1–16, 2025.

[8] Y.-L. Jin, Z.-Y. Ji, D. Zeng, and X.-P. Zhang, "VWP: An efficient DRL-based autonomous driving model," *IEEE Trans. Multim.*, vol. 26, pp. 2096–2108, 2024.

[9] D. Li, H. Deng, T. Yu, and L. Zhang, "Multi-vehicle cooperative localization using a TOA-based simulated annealing extended Kalman filter in urban canyons," *IEEE Internet Things J.*, pp. 1–14, 2025.

[10] X. Ying, C. Xiao, W. An, R. Li, X. He, B. Li, X. Cao, Z. Li, Y. Wang, M. Hu, Q. Xu, Z. Lin, M. Li, S. Zhou, L. Liu, and W. Sheng, "Visible-thermal tiny object detection: A benchmark dataset and baselines," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 6088–6096, 2025.

[11] X. Ying, L. Liu, Z. Lin, Y. Shi, Y. Wang, R. Li, X. Cao, B. Li, S. Zhou, and W. An, "Infrared small target detection in satellite videos: A new dataset and a novel recurrent feature refinement framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–18, 2025.

[12] X. Ying, L. Liu, Y. Wang, R. Li, N. Chen, Z. Lin, W. Sheng, and S. Zhou, "Mapping degeneration meets label evolution: Learning infrared small target detection with single point supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15 528–15 538.

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

[14] 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.

[15] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 4126–4135.

[16] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Exploring decision shifts in autonomous driving with attribution-guided visualization," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 3, pp. 4165–4177, 2025.

[17] M. Zemni, M. Chen, É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "OCTET: Object-aware counterfactual explanations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15 062–15 071.

[18] A. Samadi, A. Shirian, K. Koufos, K. Debattista, and M. Dianati, "SAFE: Saliency-aware counterfactual explanations for DNN-based automated driving systems," in *IEEE Int. Conf. Intell. Transp. Syst.*, 2023, pp. 5655–5662.

[19] 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.

[20] 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.

[21] T. Li, R. Shi, T. Kanai, and Q. Zhu, "Frequency-divided learning of fine-grained clothing behavior via flexible dynamic graphs," *IEEE Trans. Vis. Comput. Graphics*, vol. 31, no. 10, pp. 9166–9178, 2025.

[22] T. Li, R. Shi, Q. Zhu, and T. Kanai, "Hybrid learning with multi-scale graphs for enhanced garment deformation approximation," *Appl. Soft Comput.*, vol. 186, p. 114149, 2026.

[23] 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.

[24] 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.

[25] 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.

[26] R. Shi, T. Li, and Y. Yamaguchi, "Output-targeted baseline for neuron attribution calculation," *Image Vis. Comput.*, vol. 124, p. 104516, 2022.

[27] 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.

[28] 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.

[29] 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.

[30] R. Shi, T. Li, Y. Yamaguchi, and L. Zhang, "Attribution explanations for decision-making in deep lane-change models," *Transp. Res. Part C: Emerg. Technol.*, vol. 180, p. 105361, 2025.

[31] 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.

[32] 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.

[33] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 1135–1144.

[34] R. Niu, Q. Wang, H. Kong, Q. Xing, Y. Chang, and P. S. Yu, "Learn to explain transformer via interpretation path by reinforcement learning," *Neural Netw.*, vol. 188, p. 107496, 2025.

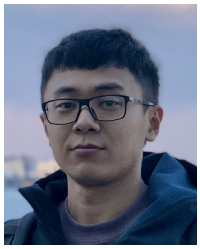
[35] R. A. Amjad, K. Liu, and B. C. Geiger, "Understanding neural networks and individual neuron importance via information-ordered cumulative ablation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7842–7852, 2022.

[36] A. Salama, N. Adly, and M. Torki, "Ablation-CAM++: Grouped recursive visual explanations for deep convolutional networks," in *IEEE Int. Conf. Image Process.*, 2022, pp. 2011–2015.

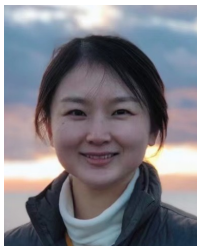
[37] Y. Peng, L. He, D. Hu, Y. Liu, L. Yang, and S. Shang, "Hierarchical dynamic masks for visual explanation of neural networks," *IEEE Trans. Multim.*, vol. 26, pp. 5311–5325, 2024.

[38] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Representations*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6034

[39] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.
[40] 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
[41] Y. Wang, H. Su, B. Zhang, and X. Hu, "Learning reliable visual saliency for model explanations," *IEEE Trans. Multim.*, vol. 22, no. 7, pp. 1796–1807, 2020.
[42] Y. Liao, Y. Gao, and W. Zhang, "Dynamic accumulated attention map for interpreting evolution of decision-making in vision transformer," *Pattern Recognit.*, vol. 165, p. 111607, 2025.
[43] T. Li, R. Shi, Q. Zhu, L. Zhang, and T. Kanai, "Spectrum-enhanced graph attention network for garment mesh deformation," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2025.
[44] 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.
[45] T. Li, R. Shi, Q. Zhu, and T. Kanai, "SwinGar: Spectrum-inspired neural dynamic deformation for free-swinging garments," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 10, pp. 6913–6927, 2024.
[46] 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.
[47] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
[48] 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.
[49] Y. Li, H. Liang, and R. Yu, "BI-CAM: Generating explanations for deep neural networks using bipolar information," *IEEE Trans. Multim.*, vol. 26, pp. 568–580, 2024.
[50] H. Wu, H. Jiang, K. Wang, Z. Tang, X. He, and L. Lin, "Improving network interpretability via explanation consistency evaluation," *IEEE Trans. Multim.*, vol. 26, pp. 11 261–11 273, 2024.
[51] C. Zhao, J. H. Hsiao, and A. B. Chan, "Gradient-based instance-specific visual explanations for object specification and object discrimination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 5967–5985, 2024.
[52] W. Ding, Y. Geng, J. Huang, H. Ju, H. Wang, and C.-T. Lin, "MGRW-Transformer: Multigranularity random walk transformer model for interpretable learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1104–1118, 2025.
[53] H. Choi, S. Jin, and K. Han, "ICEv2: Interpretability, comprehensiveness, and explainability in vision transformer," *Int. J. Comput. Vis.*, vol. 133, no. 5, pp. 2487–2504, 2025.
[54] X. Zhuang, Z. Li, C. Zhang, and H. Ma, "A cross-modal collaborative guiding network for sarcasm explanation in multi-modal multi-party dialogues," *Eng. Appl. Artif. Intell.*, vol. 142, p. 109884, 2025.
[55] D. Xue, S. Qian, and C. Xu, "Integrating neural-symbolic reasoning with variational causal inference network for explanatory visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7893–7908, 2024.
[56] M. Du, Y. Wei, Y. Tang, X. Zheng, S. Wei, and C. Ji, "St-tree with interpretability for multivariate time series classification," *Neural Netw.*, vol. 183, p. 106951, 2025.
[57] Z. Huang, S. Zhong, P. Zhou, S. Gao, M. Zitnik, and L. Lin, "A causality-aware paradigm for evaluating creativity of multimodal large language models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3830–3846, 2025.
[58] 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.
[59] A. Binder, G. Montavon, S. Lapuschkin, K. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *Int. Conf. Artif. Neural Netw.*, vol. 9887, 2016, pp. 63–71.
[60] F. R. Jafari, G. Montavon, K. Müller, and O. Eberle, "MambaLRP: Explaining selective state space sequence models," *CoRR*, vol. abs/2406.07592, 2024.
[61] 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
[62] N. Hama, M. Mase, and A. B. Owen, "Deletion and insertion tests in regression models," *Journal of Machine Learning Research*, vol. 24, no. 290, pp. 1–38, 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0560.html
[63] 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.
[64] D. Epstein, T. Park, R. Zhang, E. Shechtman, and A. A. Efros, "BlobGAN: Spatially disentangled scene representations," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13675, 2022, pp. 616–635.



Rui Shi received his Ph.D. degree in graphic and computer sciences from the University of Tokyo, Tokyo, Japan, in 2022. He is currently an associate professor 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 (Member, IEEE) 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 (Senior Member, IEEE) 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.