Exploring Decision Shifts in Autonomous Driving with Attribution-Guided Visualization

Rui Shi, Tianxing Li, Yasushi Yamaguchi, *Member, IEEE,* Liguo Zhang, *Member, IEEE,*

***Abstract*—Given the critical need for more reliable autonomous driving systems, explainability has become a key focus within the research community. In autonomous driving models, even minor perception differences can significantly influence the decision-making process, and this impact often diverges markedly from human cognition. However, understanding the specific reasons why a model decides to stop or keep forward remains a significant challenge. This paper presents an attribution-guided visualization method aimed at exploring the triggers behind decision shifts, providing clear insights into the underlying "why" and "why not" of such decisions. We propose the cumulative layer fusion attribution method that identifies the parameters most critical to decision-making. These attributions are then used to inform the visualization optimization by applying attribution-guided weights to crucial generation parameters, ensuring that decision changes are driven only by modifications to critical information. Furthermore, we develop an indirect regularization method that increases visualization quality without necessitating additional hyperparameters. Experiments on large datasets demonstrate that our method produces insightful visualization explanations and outperforms state-of-the-art methods in both qualitative and quantitative evaluations.**

***Index Terms*—Autonomous driving, visualization explanation, decision attribution, generative adversarial networks.**

I. Introduction

AUTONOMOUS vehicles have captured significant interest from the research community, owing to their potential to diminish traffic accidents and enhance transportation efficiency [1]. With the progression of deep neural network (DNN) technologies, models for autonomous driving have experienced substantial advancements. Increasingly complex DNNs are applied to real-world traffic scenarios, including trajectory tracking [2], [3], object detection [4], [5], scene understanding [6], [7], vehicle localization [8], [9], reinforcement learning-based driving control [10], [11] and motion planning [12], [13]. Notably, recent years have seen the rise of end-to-end autonomous driving models, which integrate almost all driving functionalities into DNN architectures [14].

Decision-making in autonomous vehicles is a crucial process that involves selecting a specific driving action, such as moving forward, stopping, turning left, or turning right, from a set of discrete control options. This selection is based on the current status of the ego-vehicle and its surrounding environment [15]. As the most safety-sensitive aspect of autonomous vehicle operations, the decision-making process is essential for ensuring safety and efficiency. However, the "black-box" nature of DNNs poses significant challenges for engineers attempting to understand and analyze the underlying reasons for specific decisions, particularly the causes behind shifts in decision-making.

In autonomous driving models, subtle variations in perception data can lead to different driving decisions. By intentionally generating visual modifications that trigger decision changes, we can derive meaningful insights into the decision-making process [16], [17]. Building on this idea, we introduce our optimization-based visualization method, which produces images tailored to specific decision requirements as assigned by users. This method ensures that the generated images remain similar to the original perception data yet introduce variations that significantly influence driving decisions. By comparing the generated results with the original reference image, we can identify the critical factors and conditions influencing decisions made by an autonomous driving model.

Despite the conceptual clarity of this idea, two challenges emerge during practical implementation. The first challenge involves ensuring the effectiveness of visualization explanations. Such a visualization method typically requires incorporating additional generators as prior constraints, such as generative adversarial networks (GANs) [18], and more specifically, BlobGAN [19] in our method. These constraints act on the full input information, which could be problematic in autonomous driving scenarios that feature numerous objects in one input. Optimizing the full input information often fails to accurately target the critical objects that influence decision-making, thereby introducing uncontrollable noise. To address this issue, we introduce attribution to suppress changes in unimportant regions during visualization optimization. In particular, we develop the cumulative layer fusion attribution (CLFA) method, which identifies the generator parameters most relevant to the decision. Then, we use the attributions to guide the optimization process, focusing primarily on updating critical information. This approach generates visualization explanations that closely resemble the reference while effectively highlighting the key factors influencing decision-making.

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 July 24, 2024; revised August 16, 2021.

The second challenge we address is the complexity involved in balancing image generation regularization terms. Numerous efforts have been made to design regularization methods that enhance the quality of image generation [20]–[23]. However, the introduction of regularization terms often comes with additional hyperparameters, making it difficult to achieve balance. To this end, we introduce an indirect regularization technique that can be seamlessly incorporated into the objective function without necessitating extra hyperparameters. The idea behind our method is to compel the optimization to iteratively update towards clearer and more well-defined object distinctions by applying controlled perturbations.

To validate our proposal, we conduct extensive experiments on the BDD100k [24] and BDD-OIA [25] datasets. Fig. 1 shows an example of our attribution-guided visualization explanations. Initially, the decision is "Stop," and the blob map is generated using the GAN generator. Subsequent blob-level attribution computation identifies the critical blobs. Through iterative optimization of these blob parameters under the guidance of blob attributions, we generate visualization explanations. These explanations reveal the objects that trigger a decision change from "Stop" to "Forward." The results indicate that the decision tends to shift when surrounding cars are distant; interestingly, even if the front car remains close but its brake lights are off, the decision can still shift from "Stop" to "Forward." This suggests that the autonomous driving model may be influenced by local features, such as the presence or absence of brake lights, affecting decision-making during movement and braking. With these attribution-guided visualization explanations, we are better equipped to detect these subtle biases in autonomous driving models.

The main contributions of this work are summarized as follows:

- We propose a novel visualization idea that uses attributions to constrain the generator, specifically focusing the optimization on the objects that directly impact autonomous driving decisions, thereby effectively improving the quality of visualization explanations.
- At the technical level, we develop the cumulative layer fusion attribution model by extending the Aumann-Shapley method [26], allowing for more accurate blob-level attribution computation compared to state-of-the-art alternatives. Additionally, we introduce an indirect regularization method that incorporates image processing approaches to improve visualization quality without the need for introducing extra hyperparameters.

We validate our proposal on two datasets through multiple quantitative and qualitative experiments. The results demonstrate the advantages of our method in identifying and understanding the critical information that influences driving decision-making.

II. Related Work

A. Visualization Explanation Methods

Visualization explanations involve using a predefined objective function to generate specific encodings into an image that carries a particular significance. The original concept of using

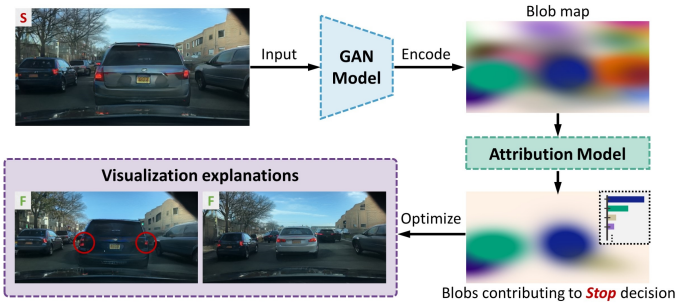


Fig. 1. Attribution-guided visualization explanation sample. Initially, the model's decision is "Stop." Through the process of encoding and attribution computation, key blobs directly related to this decision are identified, as illustrated in the bottom-right corner, where the contributions of each blob are also displayed in a bar chart. By primarily focusing on iteratively updating these identified blobs, we generate "Forward" visualization explanations, highlighting specific scene modifications that critically impact these two contrasting decisions.

an energy minimization framework for DNN visualization was developed to understand the hidden representations within networks [27]. Subsequent studies demonstrated that the methods based on StyleGAN can efficiently achieve semantically meaningful disentangled representations [28], [29]. DIEG [30] can produce diverse GAN visualizations by augmenting inverting latent embeddings with different latent samples, a technique crucial for analyzing the capability boundaries of DNNs. E4E [31] was designed to allow subsequent editing of inverted real images, producing particularly high-quality results on facial datasets. SDIC [32] developed an advanced visualization technique that incorporates spatial-contextual discrepancy information, demonstrating significant capabilities in image modification.

Some studies have applied generative models to the interpretation of autonomous driving, which include: STEEX [33], OCTET [34], and SAFE [35]. STEEX was one of the first research efforts to utilize generative models to explain autonomous driving decisions. OCTET further enhanced the diversity of visualizations by introducing BlobGAN to encode images, enabling the generation of counterfactual explanations that trigger contrast decisions. SAFE is the most recent and relevant method to our research, building upon STEEX and OCTET. It significantly improves the quality of visualizations by using saliency maps to control the optimization regions, making explanations clearer. The primary distinction between our method and these previous studies lies in the flexibility of controlling the optimized objects. While SAFE uses saliency maps to identify the regions to be updated, saliency maps struggle to cover multiple regions and often extend beyond the object itself, generating uncontrollable disturbances around the updated objects.

GAN-based visualization generation methods typically offer better generation quality, but they are inherently limited by the sample distribution of pretrained GANs, potentially leading to misleading results when facing unknown distributions. Methods that directly invert the model, although of lower generation quality, avoid the biases inherent in GANs. Yin *et al.* [20] proposed GradInversion, which balances feature representations

across all layers to execute inversion generation. Hatamizadeh *et al.* [21] introduced GradViT which enables high-quality image reconstruction for transformer-based networks. Shi *et al.* [36] designed a visualization method for attention-based visual networks, allowing the analysis of differences between different network architectures. Although these methods avoid the biases associated with pretrained GANs, their visualization qualities are subpar, making them challenging to fulfill the requirements of practical applications.

B. Attribution Explanation Methods

Attribution methods are capable of quantifying the contribution of input features to the outputs, serving as local explanation techniques. Among these methods, those leveraging backpropagation are particularly prevalent. This category includes techniques such as Gradient Shapley values [37], [38], Shapley value propagation [39], and Aumann-Shapley values [26], [40], along with other approaches that rely on gradient accumulation or Shapley value estimators [41]–[45]. Other pixel attribution methods like GradCAM [46] and LRP [47] show promise but are challenging to adapt to GAN models well with only minor modifications, leading us to exclude these techniques from our consideration in further comparison.

Generally, current studies on DNN attributions have effectively identified critical regions at the image level. Nonetheless, our work necessitates the computation of attributions at the blob level. A unique aspect is that the blob generator is not directly tied to the decision model but instead functions as an independent model. Ensuring that critical information from the decision model is not lost and the backpropagation to the blob parameters is accurately implemented are central challenges in designing a blob attribution computation model.

C. Discussion on Related Work

Several established techniques exist for visualization generation. The most relevant to our work include E4E, SDIC, STEEX, OCTET, and SAFE. Technically, E4E and SDIC have addressed the challenge of encoding images into latent representations using GANs and subsequently reconstructing the images from these representations. STEEX and OCTET established the basic framework for visualization in the autonomous driving domain. More recently, SAFE has significantly improved the visualization quality by incorporating saliency maps into the generation process, actively controlling the optimization regions. Because this optimization process leaves most pixels unchanged, it results in a substantial increase in visual similarity.

Despite the maturity of visualization techniques, existing methods often neglect object-level information crucial for decision-making. We argue that this information is essential for two key visualization requirements: maintaining similarity between the visualization and the original input, and enabling effective alteration of the original driving decision. For example, if a driving decision is primarily influenced by the taillights of the car ahead, then the ability to identify and manipulate those specific taillights is crucial for generating a meaningful visualization. Without access to this object-level information, visualizations risk being ineffective in altering the decision or may introduce excessive changes to the original input, rendering the visualization difficult to analyze. Therefore, we introduce decision attribution into visualization generation and propose cumulative layer fusion attribution (CLFA) to identify the object-level information influencing the driving decision. This attribution then serves as a guide for the visualization process. Our strategy focuses on primarily optimizing and updating the objects influencing the decision to generate the visualization. This targeted optimization facilitates decision alteration while minimizing changes to the original input.

III. Attribution Computation

In the process of visualization, the conventional practice of employing generators as a priori constraints generally involves processing the entire image. This way can hinder accurate modifications of key regions that influence decision-making. As a solution, we introduce attributions to detect key objects in traffic scenes, thereby minimizing interference from irrelevant objects. In this section, we start by discussing the widely used Aumann-Shapley attribution method. We then present our attribution method, cumulative layer fusion attribution (CLFA), specifically designed for GAN generators in autonomous driving scenarios.

A. Aumann-Shapley Attribution

A crucial aspect of our goal is to determine which objects in a given scenario most significantly influence the decision-making of an autonomous driving model. A well-established method for deriving such explanations is the Aumann-Shapley (AS) attribution technique, originally developed in the context of game theory. This technique quantitatively evaluates the contribution of each factor to the final decision. In adapting the AS method for DNNs, consider an input X and a baseline $\bar{X}$, the latter representing the absence of information. The AS method introduces a path function $\mu(t) = \bar{X} + t(X - \bar{X})$, which ensures that $\mu_i(0) = \bar{x}_i$ and $\mu_i(1) = x_i$, with $t \in [0, 1]$. The AS attribution for each feature i is then calculated by integrating the gradient of the model output along this path:

$$\phi_i = \Delta x_i \int_{t=0}^{1} \frac{\partial f^d\left(\mu(t)\right)}{\partial x_i} dt, \tag{1}$$

where f^d represents the output decision of the model f. The term $\Delta x_i = x_i - \bar{x}_i$ is the difference in feature values between the input and baseline. This integral computes how sensitively the model's output responds to changes as the input transitions from the baseline to the actual input. In practical applications, direct integration in DNNs is not feasible due to computational complexity and high dimensionality. Consequently, discrete approximations of the AS values are necessary. We use the Gauss-Legendre quadrature for this approximation:

$$\phi_i = \Delta x_i \sum_{k=1}^{K} \frac{1}{\left(1 - \xi_k^2\right)\left[P_K'(\xi_k)\right]^2} \frac{\partial f^d\left(\mu(\xi_k)\right)}{\partial x_i}, \tag{2}$$

where K is the number of discrete points used in the approximation, ξ_k denotes the chosen quadrature points within the

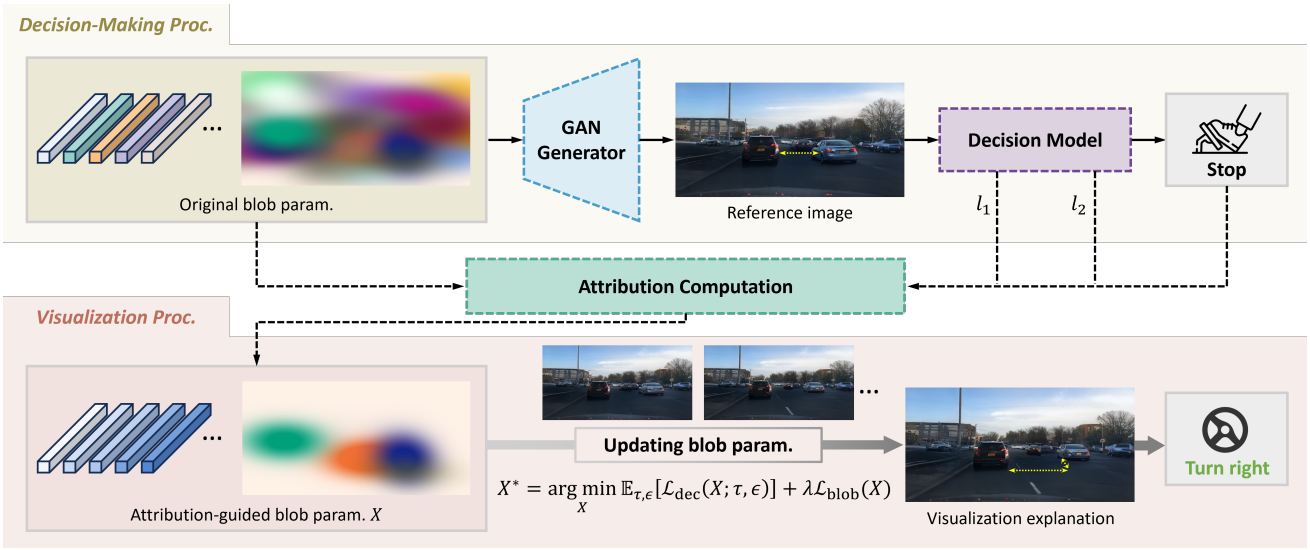


Fig. 2. An overview of the visualization explanation generation. The decision process information primarily aids in the attribution computation to identify the key blobs. Subsequently, by optimizing the visualization objective function, we can progressively derive the visualization explanation that results in a change in the decision. In this scenario, the decision changes from "Stop" to "Turn Right" when the car positioned to the right is farther away. Note the change in the length of the yellow arrows in the images. Typically, modifying only the critical information directly related to the decision can lead to a significant shift in the decision.

standard interval for integration, and $P'_K(\xi_k)$ is the derivative of a polynomial at the point ξ_k. This equation quantifies the contribution of individual input features to the output.

B. Cumulative Layer Fusion Attribution

In the preceding section, we introduce the AS method. However, our specific application focuses on attributing contributions not to the images themselves but to the decision-relevant parameters within a generator. Blob-level attributions allow us to directly modify these parameters, thereby influencing driving decisions. The decision-making process involves two independent models: the GAN generator and the autonomous driving decision model, as shown in the upper part of Fig. 2.

To this end, we extend the concepts of X and its baseline zero $\bar{X}$ to denote the blob parameters in the generator, while $f^d(X)$ represents the output decision based on these input parameters. Regarding the generator, we employ a differentiable BlobGAN due to its capability to encode objects into semantic blobs. This feature allows us to adapt the original scenario by modifying the blob parameters. Each blob is characterized as an ellipse defined by several parameters: the centroid, scale, aspect ratio, and rotation angle. Additionally, each blob is assigned structural and stylistic features ψ, allowing for detailed modifications such as altering the shape and color of traffic lights within a traffic scene.

Direct computation of the gradient of the BlobGAN parameter path function with respect to the decision output often results in information loss, leading to blob attributions that may not accurately represent the specified decisions. A similar issue has been noted in previous research [46], [48], which suggests that incorporating information from other hidden layers can alleviate these inaccuracies. Inspired by these insights, we propose cumulative layer fusion attribution (CLFA), a method that accumulates additional hidden layers into the AS computational model to achieve more accurate and focused attributions.

First, the input parameters X are processed through l layers, represented by the function of $f^{[l]}$, to obtain the intermediate feature map:

$$A^{[l]} = f^{[l]}(X). \tag{3}$$

Based on the feature map $A^{[l]}$, its baseline $\bar{A}^{[l]}$ can be derived by minimizing its spatial dimensions of $A^{[l]}$, followed by broadcasting the minimized values across these dimensions. Next, the interpolant $\tilde{A}^{[l]}$ along the path $\mu^{[l]}$ from $\bar{A}^{[l]}$ to $A^{[l]}$ can be expressed as:

$$\tilde{A}^{[l]} = \mu^{[l]}(t^{[l]}). \tag{4}$$

Finally, we define the cumulative layer fusion attribution by integrating contributions over intermediate states from multiple layers with nesting integrals:

$$\begin{aligned}\phi_i = \Delta x_i \int_{t=0}^{1} \frac{\partial f^d\left(\mu^{[0]}(t^{[0]})\right)}{\partial x_i} \cdots \\ \int_{t=0}^{1} \frac{\partial f^d\left(\mu^{[l]}(t^{[l]})\right)}{\partial x_i} dt^{[l]} \cdots dt^{[0]}.\end{aligned} \tag{5}$$

For the sake of notational simplicity, we define f^d as a function that changes based on its input. For instance, if the input is $\mu^{[0]}(t^{[0]})$, then f^d represents the model that takes blob parameters as input and outputs a decision. If the input is the l_{th} layer's feature $\mu^{[l]}(t^{[l]})$, then f^d refers to a truncated network model at this layer, where the input consists of features from the l_{th} layer and the output is the decision. In practice, to implement the CLFA method, we perform a discrete numerical approximation following the same principles as Eq. (2). Compared to the original AS method, the advantage of CLFA lies in its ability to accumulate contributions from various intermediate states throughout the

propagation. This accumulation enhances both the stability and the magnitude of the gradient signal across deep layers, offering a more detailed and comprehensive understanding of how input parameters across different layers influence the final decision.

IV. Attribution-Guided Visualization Explanation

Visualization has the capacity to create various scenarios that trigger different driving decisions, and this process can be formulated as an optimization problem. For our visualization explanations, we aim to generate an image that closely resembles the reference image, with only key objects modified. These modifications are designed to provoke shifts in decision-making, allowing us to understand autonomous driving decisions by observing the differences between the reference and the visualization images. An overview of our attribution-guided visualization process is shown in Fig. 2. Note that, the visualization process in the lower part of Fig. 2 also incorporates the GAN generator and the decision model for optimization generation.

The principle behind the visualization explanation is simple; however, the key challenge lies in generating high-quality results. We find that visualization processes tend to favor low-frequency information, which often results in blurred outputs. Although numerous regularization techniques have been developed to mitigate this issue [20]–[23], another challenge remains: the difficulty in balancing the hyperparameters associated with each additional regularization term.

To address this, we develop an indirect regularization method. This method introduces random perturbations during the optimization, indirectly promoting the generation of clearer images to compensate for these disturbances. Specifically, we incorporate two differentiable image processing techniques, including random jitter and generalized Kuwahara filter [49]. We find that edge-preserving filters generally yield satisfactory results. Ultimately, we select the Kuwahara filter for its superior visual effects on high-contrast images. Our indirect regularization can be represented as:

$$f_{\text{reg}}(X;\tau,\epsilon) = f_{\text{jit}}\left(f_{\text{kuwa}}\left(I(X);\tau\right);\epsilon\right), \tag{6}$$

where the image I is generated with the blob parameters X. This sequence involves first applying a filter with a randomly selected weighting function from eight sectors of a disc as discussed in [49]. The weighting function is indexed with a discrete random variable τ which is uniformly distributed over the set $\{1,\cdots,8\}$. Additionally, the f_{jit} function introduces a controlled amount of spatial jitter, parameterized by ϵ, ranging from $-T$ to T. Here, T can be empirically set as the largest integer less than $(H+W)/100$. These regularization disturbances can enhance the overall robustness and visual coherence of visualization images.

With these foundational concepts established, we can formally define our attribution-guided visualization objective function, which comprises two distinct loss functions. The decision loss is specifically designed to influence and alter the driving decision. The blob loss is guided by attributions and aims to maintain the stability of blob parameters that are not directly related to the decision-making process. The objective function can be represented as:

$$X^* = \arg\min_X \mathbb{E}_{\tau,\epsilon}\left[\mathcal{L}_{\text{dec}}(X;\tau,\epsilon)\right] + \mathcal{L}_{\text{blob}}(X), \tag{7}$$

where $\mathbb{E}_{\tau,\epsilon}[\cdot]$ indicates the expectation of two random variables τ and ϵ. The hyperparameter λ is used to balance these two loss terms, ensuring that they have comparable magnitudes. For the decision term, we follow the visualization framework discussed in [50] to implement our loss by pushing the decision opposite to the original prediction:

$$\mathcal{L}_{\text{dec}}(X) = (s-1)\log\left(1-f^d(X)\right) - s\log\left(f^d(X)\right), \tag{8}$$

where $s \in \{0.05, 0.95\}$ represents the score opposite to the original decision. We do not use a binary s for stable numerical optimization. Decision losses can be accumulated multiple times. For instance, to shift to a specific decision, we simply need to incorporate an additional decision loss, *i.e.*, the decision index d does not necessarily correspond to the decision output by the autonomous driving model.

To ensure the visualization results remain visually similar to the original image, we define a set $\mathbb{U}$, which includes indices corresponding to blobs with low attributions. These indices reflect blobs that are less significant for the model's decision-making process. The sum of the attributions of the blobs within the set does not exceed 40% of the total sum. The set $\bar{\mathbb{U}}$ comprises the indices of the rest of the blobs with higher contributions. The corresponding loss function, aimed at minimizing the differences in unimportant regions, is expressed as follows:

$$\mathcal{L}_{\text{blob}}(X) = \lambda \sum_{b\in\mathbb{U}} \|X^b_{\text{ori}} - X^b\|_2 + \sum_{b\in\bar{\mathbb{U}}} e^{-(\gamma^b-\beta)} \|X^b_{\text{ori}} - X^b\|_2, \tag{9}$$

where X^b_{ori} and X^b represent the blob parameters of the reference and the visualization explanation, respectively. λ is a hyperparameter to balance two terms. $\gamma^b = \sum_i \phi^b_i$ is used to weight the important blob parameter changes. Upon sorting the blob attributions, the critical blob index b_1 is determined when the cumulative attribution score reaches a predefined threshold of the total attribution score, *e.g.*, 60%. The subsequent blob index is denoted as b_2. β represents the average value between γ^{b_1} and γ^{b_2}. In our method, we specifically use the attributions of blob features ψ to define the set $\mathbb{U}$, as these features contain more detailed information about the object semantics compared to other parameters of the blob ellipses. To mitigate the effects of outlier values and ensure a stable optimization process, we maintain an exponential decay weight greater than 0.01. The optimized result X^* is forwarded into the GAN generator to produce the final visualization explanation.

V. Experimental Results and Discussions

A. Implementation Details

Our experiments utilize the BDD100k dataset, which consists of 100k images. Additionally, we incorporate the BDD-OIA, an extension of 20k scenes from BDD100k, annotated with binary labels for potential ego-vehicle actions: "Forward," "Stop," "Turn Left," and "Turn Right." We train a multi-label

binary DenseNet on the BDD-OIA dataset as the model for autonomous driving explanations, following the implementation previously described in [51]. We also train a BlobGAN on the BDD100k dataset as per the settings [19]. The number of blobs is increased to 40 to match the object classes of panoptic segmentation labels in the BDD dataset. Before generating an explanation, we need to obtain the blob representation of an image. This is achieved through an inversion process, which we implement according to the technique described in [19].

Based on our experimental observations and attribution computational performance, we empirically select three specific layers for attribution computation. These layers include the input layer, the layer after two max pooling operations, and the penultimate max pooling layer. The input layer represents the first layer that processes blob parameters and is located in the GAN model instead of the decision model.

We set the sample points $K = 20$ for the input layer and $K = 10$ for other layers. This configuration is sufficient for the attributions to nearly sum up to the specific output score for our case, *i.e.*, roughly satisfying the "completeness" axiom of Shapley values as stated in [52]. Additionally, our attribution computation process is nested. In particular, if we select the layers $\{0, l_1, l_2\}$, then, based on Gauss-Legendre quadrature, the definition of the interpolant $\tilde{A}$ is defined as follows:

$$\begin{cases} \tilde{X} = \frac{1}{2}\left((1-\xi_{k^{[0]}})\bar{X} + (1+\xi_{k^{[0]}})X\right), \\ \tilde{A}^{[l_1]} = \frac{1}{2}\left((1-\xi_{k^{[l_1]}})\bar{A}^{[l_1]} + (1+\xi_{k^{[l_1]}})A^{[l_1]}\right), \\ A^{[l_1]} = f^{[l_1]}\left(\mu^{[0]}(\xi_{k^{[0]}})\right), \\ \tilde{A}^{[l_2]} = \frac{1}{2}\left((1-\xi_{k^{[l_2]}})\bar{A}^{[l_2]} + (1+\xi_{k^{[l_2]}})A^{[l_2]}\right), \\ A^{[l_2]} = f^{[l_2]}\left(\mu^{[l_1]}(\xi_{k^{[l_1]}})\right). \end{cases} \tag{10}$$

Although the attribution computation is performed on the full blob parameters, for the purpose of key blob detection, specifically determining the set $\mathbb{U}$ in Eq. (9), we sum attributions of features ψ as a scalar value for each blob.

For the visualization explanation optimization, we set the hyperparameter λ to 10 to ensure that most irrelevant image regions do not change obviously. We employ the Adamax optimization algorithm [53] and implement an exponential decay of the learning rate, which is adjusted in response to changes in the loss.

B. Evaluation of Blob Attribution

In this section, we assess our attribution method against various state-of-the-art methods. Attribution computation is crucial for our visualization explanation task, aiming to precisely identify information significantly influencing the decision-making process. To validate the accuracy of blob-level attributions, we conduct a comparative analysis against several methods including GradShap [38], IDGI [41], AS [40], INA [42], and PropShap [39].

In Fig. 3, we present three sets of examples for comparison. The first column displays the original inputs alongside their corresponding blob maps. The first row of each example shows the critical blobs determined by blob-level attributions. We sort the blobs according to their attributions and identify those whose combined contributions account for 60% of the total attribution sum. That is, only the highly relevant blobs are depicted. Intuitively, the critical blobs identified by our method align closely with human cognition.

TABLE I
QUANTITATIVE COMPARISON OF DECISION SCORE CHANGES AFTER COLORING GRAY NON-ESSENTIAL REGIONS.

	Foward	Stop	Turn Left	Turn Right
GradShap	0.278	0.273	0.331	0.359
IDGI	0.220	0.214	0.275	0.277
AS	0.251	0.256	0.312	0.322
INA	0.304	0.298	0.354	0.376
PropShap	0.237	0.243	0.293	0.306
CLFA	**0.181**	**0.178**	**0.247**	**0.261**

To quantitatively validate our method, we design an indirect evaluation metric for blob attribution. We first remove the most important blobs from the original blob maps, ensuring that no more than five blobs are removed, to detect significant regions in driving scene images. Then, we generate a new image with only unimportant blobs and identify unchanged regions compared to the reference image. These regions of lesser importance are then colored gray as shown in Fig. 3. Previous methods show some failures in targeting these critical regions or included irrelevant objects, leading to redundancy. In contrast, our method accurately identifies critical regions, such as the brake lights in the second example and the road in the last example, with minimal impact from irrelevant objects. We then forward these images, where non-critical regions have been de-emphasized, into the decision model to obtain new decision scores, denoted as s_{new}. The change ratio in score is quantified by $|s_{\text{new}} - s_{\text{ori}}|/s_{\text{ori}}$.

The quantitative results of the different methods for four driving decisions are presented in Table I, where a lower value indicates better performance. Our CLFA consistently demonstrates the smallest variation in score changes across various decision scenarios, highlighting its robustness in blob attribution.

We also note that decisions related to "Forward" and "Stop" show less variability compared to those involving "Left" or "Right" turns. This consistency is likely due to the reliable identification of influential objects such as tail lights and traffic lights, which are crucial for "Forward" and "Stop" actions. These elements are consistently highlighted by various methods, ensuring stable decisions even when other regions are de-emphasized. Conversely, decisions about turning left or right tend to be more affected by variations in the surrounding environment, making them more prone to changes when subjected to similar perturbations.

C. Evaluation of Visualization Results

In this section, we explore and compare various visualization explanation methods. Our analysis is not limited to GAN-based approaches, including DGN-AM [27], SDIC [32], DIEG [30], OCTET [34], SAFE [35], and E4E [31], but also

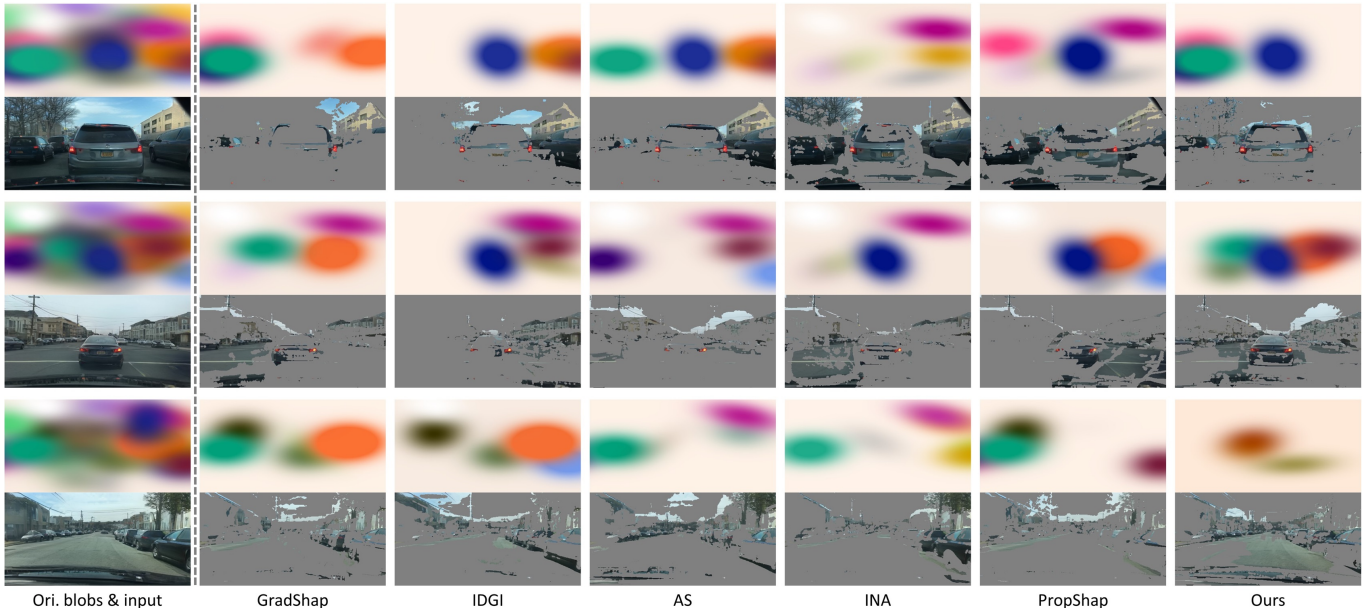


Fig. 3. Qualitative comparison of blob-level attribution. We show three sets of examples, the first and second of which are "Stop" decisions and the third is a "Forward" decision. For each example, the first row displays the critical decision-relevant blobs identified by various attribution methods. The second row illustrates the critical regions impacting the decision, highlighted after coloring gray the less relevant regions.

encompass direct image space optimization approaches such as GradViT [21], GradInv [20], and VisComp [36].

Among the GAN-based methods, DGN-AM stands out as a pioneering approach, being one of the first to leverage GANs for diverse explanation experiments. Unlike newer methods that employ complex structures like StyleGAN [28], DGN-AM utilizes a relatively straightforward convolutional neural network architecture. Given that BlobGAN also bases its architecture on StyleGAN, we can seamlessly substitute the GAN models in SDIC, DIEG, and E4E with our trained BlobGAN. OCTET and SAFE, originally developed for BDD datasets, can be directly compared with our method. Some other visualization methods are designed to understand neural network features (*e.g.*, VisComp and DGN-AM) or to reconstruct images from specific feature representations (*e.g.*, GradViT and GradInv). While these methods are not directly applicable to our task, their optimization regularization techniques prove useful for achieving high-quality visualization results. We incorporate these techniques into our experimental framework for comparison. To enable fair comparison in autonomous driving scenarios, we use the same decision loss function as our method for all compared methods. Furthermore, we conduct an experiment to evaluate the effects of removing optimization regularization and attribution guidance, denoted as "w/o Reg" and "w/o Attr," respectively.

Quantitative evaluations in our study still focus on two primary metrics: similarity and effectiveness. We use the one minus the Learned Perceptual Image Patch Similarity (LPIPS) score [54] to measure the perceptual distance between the visualization results and the original inputs. Similar scenarios leading to different decisions can better assist in examining hidden biases in autonomous driving models. Effectiveness, on the other hand, is determined by the success rate at which visualization explanations change decisions. Fig. 4 illustrates

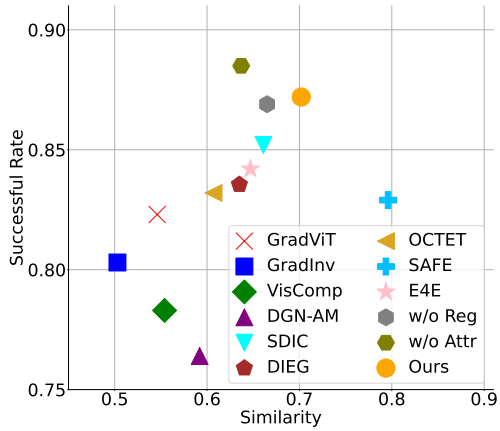


Fig. 4. Similarity (x-axis) against the rate of successful decision changing (y-axis) for all compared methods. The most valuable visualization explanation should be in the top-right corner.

that our method not only achieves a high similarity score but also effectively alters the original decisions. Eliminating attribution guidance (*i.e.*, "w/o Attr") slightly improves success rates but adversely affects similarity. The absence of indirect regularization (denoted as "w/o Reg") leads to more visually disordered traffic scenes, further decreasing similarity scores.

When comparing with other GAN-based methods, we observe significant variations in performance across different approaches. For example, while the latest discrepancy map learning hourglass module and attention fusion techniques exhibit strong performance on facial datasets [32], they fall short in visualizing traffic scenes, as indicated by the light blue inverted triangle in Fig. 4. Although SAFE can achieve the highest similarity, its approach of keeping pixel-level changes minimal makes it difficult to alter decisions, thereby impacting

TABLE II
QUANTITATIVE COMPARISON WITH DIFFERENT VISUALIZATION METHODS.

	FID ↓	KID ↓	1-LPIPS ↑	SSIM ↑	Time ↓
GradViT	58.385	0.085	0.546 (±0.293)	0.474 (±0.327)	4.93 (±0.31)
GradInv	64.392	0.089	0.503 (±0.367)	0.488 (±0.314)	**4.75** (±0.42)
VisComp	56.227	0.077	0.554 (±0.307)	0.516 (±0.297)	5.16 (±0.35)
DGN-AM	48.793	0.073	0.592 (±0.301)	0.542 (±0.294)	6.58 (±0.57)
SDIC	31.691	0.054	0.661 (±0.247)	0.616 (±0.245)	8.42 (±0.83)
DIEG	34.425	0.059	0.635 (±0.251)	0.588 (±0.284)	8.14 (±0.74)
OCTET	54.951	0.061	0.607 (±0.302)	0.571 (±0.281)	7.53 (±0.72)
SAFE	28.453	0.059	**0.796** (±0.124)	0.599 (±0.262)	6.74 (±1.13)
E4E	34.310	0.053	0.647 (±0.271)	0.591 (±0.274)	7.15 (±0.98)
w/o Reg	35.048	0.061	0.665 (±0.249)	0.593 (±0.283)	2.16+6.67 (±1.02)
w/o Attr	37.353	0.062	0.637 (±0.274)	0.584 (±0.291)	6.81 (±0.87)
Ours	**24.767**	**0.048**	0.702 (±0.224)	**0.642** (±0.236)	2.16+6.83 (±0.92)

Fig. 5. Visualization results generated by different methods. Our results consistently exhibit a high similarity and effectively change decisions by modifying small yet critical details, reflecting the key information that contributes to decision shifts in the scenario. These images are best viewed on screen.

the effectiveness of visualization explanations. Moreover, direct image-space generation methods (*i.e.*, GradViT, GradInv, and VisComp) yield lower-quality images, all scoring below 0.6 in terms of similarity. This result shows the necessity of utilizing sophisticated GAN architectures for our visualization tasks.

The evaluation of similarity is multifaceted, prompting us to select various metrics to capture different aspects of image quality and realism. We use the LPIPS metric to measure perceptual distance. Additionally, we integrate the Fréchet Inception Distance (FID) [55] and Kernel Inception Distance (KID) [56] to evaluate the statistical similarity and diversity of the generated images compared to the inputs, and use the Multi-Scale Structural Similarity Index (MS-SSIM) [57] for direct comparisons in the image space. For non-statistical metrics, we add their numerical ranges for reference. Table II shows that our method outperforms other state-of-the-art methods across most metrics. For SAFE, we mainly refer to the data provided in the original paper [35]. We find that our attribution guidance is crucial for maintaining the proximity of visualization explanations. Directly updating all parameters without specific guidance (*i.e.*, w/o Attr) leads to a significant reduction in similarity, highlighting the importance of targeted interventions in the visualization process. Among the comparison methods, SDIC, SAFE, and E4E are notable for generating the most interpretable results.

The final column of Table II presents the visualization generation time for each method, allowing for an efficiency comparison. Our method needs an additional overhead of approximately 2 seconds due to the attribution computation. During the visualization optimization process, most GAN-based generation methods exhibit similar time consumption, typically requiring 6-8 seconds depending on scene complexity and optimization objectives. Visualization methods that do

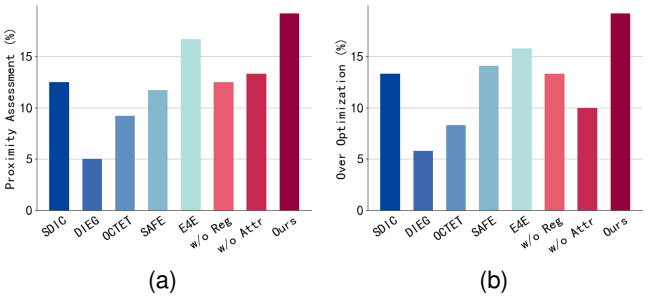


Fig. 6. User study results of visualization explanations.

TABLE III
QUANTITATIVE COMPARISON OF DECISION SCORE CHANGES USING DIFFERENT LAYER SELECTIONS.

	Foward	Stop	Turn Left	Turn Right
l_l	0.288	0.292	0.347	0.351
l_m	0.283	0.289	0.343	0.346
l_{h1}	0.263	0.267	0.328	0.319
l_{h2}	0.262	0.268	0.332	0.322
(l_l, l_m)	0.235	0.239	0.307	0.291
(l_l, l_{h2})	0.196	0.189	0.254	0.273
(l_m, l_{h1})	0.219	0.218	0.269	0.281
(l_m, l_{h2})	0.221	0.217	0.282	0.288
Ours (l_l, l_{h1})	**0.181**	**0.178**	**0.247**	**0.261**

not rely on GANs, such as GradViT, GradInv, and VisComp, offer faster generation speeds, typically around 5 seconds. All reported times are based on single-batch processing using a computer equipped with an I9 13900K CPU and an NVIDIA GeForce RTX 4090 GPU. Due to the computational complexity of visualization generation, current techniques are not capable of real-time explanation generation and are therefore best suited for offline interpretation of model decisions.

A qualitative visual comparison of our method against the three top-performing methods is presented in Fig. 5. Our visualizations not only maintain high fidelity to the reference images but also offer insightful interpretations of decision changes within the autonomous driving model. For instance, in the first column of Fig. 5, where the autonomous driving decision shifts from "Forward" to "Turn Left," our visualization does not simply remove the leading vehicle; it reveals a navigable path to the left. This nuanced representation facilitates a more in-depth analysis of the scenarios that might induce a change in the model decision. In the second column, our method triggers a decision shift by repositioning the leading vehicle while best preserving its shape. This careful balance between inducing change and maintaining visual consistency is crucial for generating meaningful and interpretable visualizations. Both quantitative and qualitative results confirm that our method is strong in exploring decision shifts of autonomous driving models.

D. Visualization Evaluation User Study

Although we have used quantitative metrics to benchmark our method against existing methods Sec. V-C, visualization, as shown in Fig. 5, is inherently subjective. Therefore, quantitative experiments alone are insufficient for a comprehensive validation of the generated explanations. To address this, we design two progressive user studies to evaluate the effectiveness of our visualization explanations in inducing decision changes while preserving fidelity to the original input. These studies involved 60 participants recruited from three universities and a community. Of these participants, 22 are actively involved in autonomous driving research, with 9 possessing practical driving experience. The remaining 38 participants have limited familiarity with autonomous driving, and 17 of them have driving experience.

Proximity assessment. Participants were asked to examine 20 randomly selected sets of images. Each set presented visualizations generated by different methods in random order, alongside the original image and its corresponding driving decision displayed consistently. Participants selected the three visualizations they perceived as most closely resembling the original input. The top three methods were assigned scores of 3, 2, and 1, respectively, while all other methods received a score of 0. Human judgment of image similarity often diverges significantly from numerical metrics. People tend to assess similarity based on overall style and structural layout. For example, SAFE produces hallucination-like artifacts in certain samples, which are readily perceptible to the human eyes but may not significantly impact quantitative metrics. Conversely, E4E's superior preservation of overall structure correlates with its stronger performance in the user study, as shown in Fig. 6a. Generally, user study results provide a more accurate reflection of human perception and understanding of the visualization explanations. For this indicator, our method outperforms all other methods, demonstrating that our visualizations are better aligned with human perception and facilitate human understanding of model decision-making.

Over optimization evaluation. We further presented participants with the decisions resulting from the visualizations and asked them to evaluate which visualization best helped them locate the key information responsible for the decision shift, again allowing them to select top three methods. This experiment requires the visualization explanations to effectively highlight the updated regions, thereby aiding user understanding of the reasons behind the decision change. A higher selection percentage indicates better comprehensibility. As shown in Fig. 6b, our method also achieved the best performance.

E. Ablation Study of Layer Selection

In this section, we evaluate the impact of layer selection within CLFA computation, using both attribution and visualization metrics. We use different layers: for the lower layer, we use the layer post two maxpoolings, denoted as l_l; for the middle layer, we choose the middle maxpooling layer l_m; and for the higher layers, we select the second to last layer l_{h1} and the fourth to last layer l_{h2}. To validate our layer selection strategy, we perform attribution calculations using both individual layers and pairs of layers. Then, we use the change ratio score as stated in Sec. V-B to evaluation

TABLE IV
QUANTITATIVE COMPARISON OF DIFFERENT SIMILARITY METRICS USING VISUALIZATIONS GENERATED WITH DIFFERENT LAYER SELECTIONS.

	FID ↓	KID ↓	1-LPIPS ↑	SSIM ↑
l_l	29.339	0.056	0.682	0.601
l_m	30.392	0.054	0.688	0.612
l_{h1}	28.547	0.053	0.684	0.611
l_{h2}	28.367	0.053	0.687	0.607
(l_l, l_m)	26.628	0.051	0.691	0.627
(l_l, l_{h2})	25.421	0.048	0.699	0.631
(l_m, l_{h1})	26.397	0.049	0.692	0.634
(l_m, l_{h2})	25.792	0.049	0.694	0.629
Ours (l_l, l_{h1})	**24.767**	**0.048**	**0.702**	**0.642**

these attributions. The results shown in Table III demonstrate that single-layer selection struggles to effectively accumulate features, generally leading to suboptimal performance. Utilizing layer pairs, however, significantly improves performance. Our final choice is (l_l, l_{h1}) as detailed in Sec. V-A. This choice is based on the established understanding that low-level features capture fundamental scene information, while high-level features possess greater discriminatory power [36]. By combining these layers, we leverage the strengths of both low-level and high-level information, and the experiments confirm that this integration effectively enhances the accuracy of the attribution calculations.

We further validate our layer selection by generating visualizations using attributions calculated from the different layer combinations and subsequently evaluating their quality, as presented in Table IV. Although the performance differences across the various layer selections are not substantial, our chosen layer pair consistently yields the most effective results, enabling more precise localization of the key information contributing to the driving decision.

F. Ablation Study of Blob Ratios

In this section, we discuss the blob ratio setting for determining critical blobs. After the computation of attributions, we select a specific proportion of blobs, whose combined attributions account for more than 60% of the total sum, as discussed in Sec. V-B. This selection forms the basis for our subsequent visualization optimization.

To enhance the interpretability of visualization explanations for the decision-making process, it is important that the explanations should exhibit two key properties: similarity and effectiveness [50]. Similarity ensures that the explanation results closely resemble the inputs, thereby accurately revealing the critical information that influences decisions and uncovering potential model deficiencies. Effectiveness ensures that the explanation results have a tangible impact on the driving decisions. For example, if the optimization goal is set to "Stop," then the decision for the explanation image should shift from the original decision to "Stop."

Specifically, the similarity of the explanation image to the reference is quantified by using one minus LPIPS score [54], and the effectiveness is measured by the change ratio of the decision score values. Our objective is to achieve both high

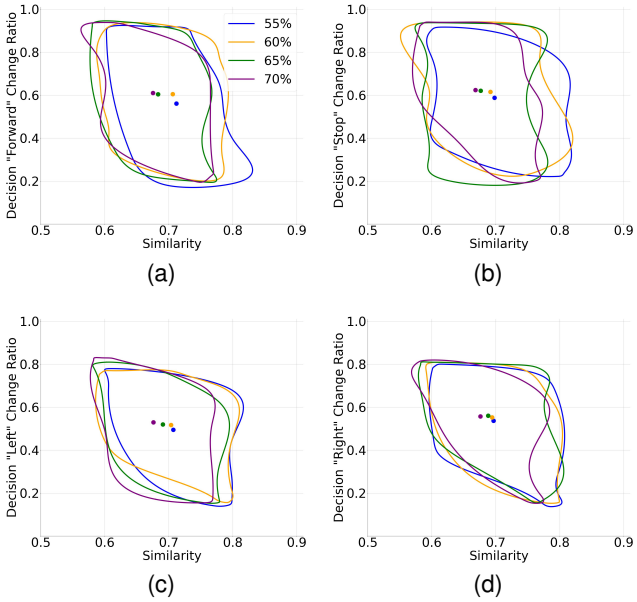


Fig. 7. Evaluation of different blob ratios. We plot image similarity (x-axis) against the ratio of decision score change (y-axis) for four distinct blob ratios under four types of decisions: (a) Forward, (b) Stop, (c) Turn Left, and (d) Turn Right.

similarity and substantial decision score changes. As shown in Fig. 7, the most insightful explanations should be positioned in the upper-right part. The closed curves are formed by connecting the data points.

We conduct a sensitivity analysis to determine the optimal blob ratio for generating effective visualizations. The experimental results indicate that a balance between similarity and effectiveness in altering the decision is best achieved when the blob ratio is set at 60%, particularly noticeable in "Forward" and "Stop" decisions. These decisions appear more susceptible to alteration through visualization explanations. We hypothesize that this sensitivity may stem from an imbalance in the distribution of decision labels within the BDD dataset, which contains significantly more "Forward" and "Stop" examples. Furthermore, "Forward" and "Stop" decisions often correlate with more readily discernible objects, such as the distance to the preceding vehicle, the state of brake lights, and traffic signals, which directly influence the decision. Conversely, decisions involving turning left or right typically entail more complex scene changes, making it more challenging to generate effective visualizations that both maintain fidelity to the original input and induce a decision change.

To further explore the impact of the blob ratio, we change this parameter across a range of values (*i.e.*, 55%, 60%, 65%, and 70%) and evaluate the resulting visualizations in terms of both similarity and effectiveness. As shown in Fig. 7, the results confirm that a 60% blob ratio consistently provides the best trade-off between these two competing metrics across different decision types. Higher blob ratios tend to prioritize similarity at the expense of effectiveness, while lower ratios prioritize effectiveness. Based on these findings, we adopted a 60% blob ratio for our experiments.

TABLE V
QUANTITATIVE COMPARISON OF SIMILARITY ACROSS DIFFERENT REGULARIZATIONS.

	FID ↓	KID ↓	1-LPIPS ↑	SSIM ↑
w/ TV	27.483	0.055	0.672	0.608
w/ BN	33.562	0.058	0.636	0.592
w/ TV & BN	26.358	0.054	0.679	0.611
w/ jit	29.427	0.055	0.682	0.627
w/ Kuwa-5	29.103	0.055	0.687	0.628
w/ Kuwa-7	27.824	0.054	0.691	0.633
w/ Kuwa-9	28.628	0.054	0.689	0.630
w/ jit & Kuwa (Ours)	**24.767**	**0.048**	**0.702**	**0.642**

G. Ablation Study of Regularization Techniques

In this section, we compare our regularization method with a set of classic image generation regularization techniques, namely total variation (TV) and batch normalization (BN) regularization. Such a combination has become standard in many visualization methods [20], [21], [58]. Compared to these techniques, the main advantage of our proposal is its capacity to integrate various image processing methods without introducing additional hyperparameters. This advantage is particularly useful when there are large variations in the range of the objective function, as it avoids the need to balance parameters across multiple loss terms. Furthermore, we use image quality assessment metrics to compare different regularization methods, as shown in Table V.

As an ablation reference, we use only one image processing regularization in our method, denoted as "w/ jit" and "w/ Kuwa." We also test various sizes for Kuwahara filter (*i.e.*, 5, 7, and 9) and determine that a size of 7 provides optimal performance for our model. This setting is used as the default in our experiments. It is important to note, however, that this filter size should be adjusted based on the input size of the model to achieve the best results. Jitter processing, which introduces random perturbations during optimization, primarily impacts image sharpness. The Kuwahara filter, on the other hand, primarily affects image texture. Applying jitter in isolation leads to a more noticeable decrease in image quality compared to using the Kuwahara filter alone. The experimental results demonstrate that combining both regularization techniques yields the highest quality visualizations.

H. Bias Identification via Visualization Explanations

In this section, we use visualization explanations to reveal biases within the autonomous driving model. Previous experiments have shown that our visualizations satisfy the metrics of the explanatory task. Leveraging the high-quality results obtained, we delve deeper into evaluating how effectively these explanations can identify biases that trigger decision shifts within the model. To this end, we conducted a user study involving 50 participants recruited from three universities. Of these, 19 are actively engaged in autonomous driving research, and 9 of them have practical driving experience. The other 31 participants, less familiar with autonomous driving, include 11 who have driving experience. This varied mix of participants provides a broad perspective, helping us better understand how different levels of familiarity with autonomous driving influence decision-making perceptions.

Participants were tasked with reviewing approximately 100 randomly generated pairs of explanations, each composed of two images corresponding to different decisions, with examples shown in Fig. 8 and 9. The original inputs are shown above, and the visualizations that change the decisions are shown below. Participants were asked to pinpoint which explanations seemed inconsistent with their intuition and to explain their reasoning.

Through this interactive format, a significant number of participants highlighted two notable biases in the model trained on the BDD dataset. First, variations in tail lights often change the model's decision. As shown in Fig. 8, when the tail lights of the car ahead are off, the model typically opts to continue moving forward, disregarding the broader context. In contrast, if the tail lights are on, the model tends to decide to stop. From the feedback collected, we analyze how this dependence on the state of the tail lights affects the quality and safety of decisions made by the model. Primarily, if the model makes a braking decision solely on the tail lights' status, it might overlook other critical environmental and contextual information such as the actual speed of the car ahead, the flow of traffic, and road conditions. This over-reliance can lead to delayed responses in urgent scenarios where tail lights are not activated, or to exaggerated reactions during frequent but unnecessary tail light flashes. Moreover, this decision-making process may prove inadequate under conditions where tail lights are non-standard, malfunctioning, or not clearly visible due to unusual lighting, potentially causing inaccurate system responses and increasing accident risks.

The second bias identified in the autonomous driving model is about traffic lights. Participants observed that the model is frequently influenced by objects resembling traffic signals during its decision-making process. As shown in Fig. 9, even minor alterations in the scene, such as the emergence of a traffic light or a similar-looking light source, can directly change the model's decision. While this high sensitivity to traffic lights may appear beneficial, *e.g.*, prompting the vehicle to stop when necessary, it also introduces uncontrollable risks, particularly in complex urban scenes. For example, if the light reflecting off a building's glass facade under certain conditions mimics traffic signal colors, or if certain streetlights mirror the brightness and color of traffic signals, the model could erroneously interpret these reflections or light sources as traffic signals and decide to stop.

Our visualization explanations serve as a tool that enables developers and users to identify such biases within the model, thereby enhancing the predictability and consistency of its decisions.

VI. CONCLUSION

In this study, we introduce a novel attribution-guided optimization visualization approach designed to explore the reasons behind decision shifts in autonomous driving models. Our method focuses on identifying and updating only the key influential objects relevant to the decision, which facilitates

Fig. 8. Samples of decision-making bias triggered by tail lights.

Fig. 9. Samples of decision-making bias triggered by traffic lights.

Fig. 10. Visualizations generated with samples from the KITTI and HDD datasets.

the creation of high-quality, decision-inspired visualizations. This achievement is enabled by our proposed cumulative layer fusion attribution method and an indirect regularization technique. Our approach has demonstrated superior results compared to previous works across various evaluations. We believe that this work helps to facilitate the exploration of the decision-making process and enhance the analysis of autonomous driving systems in real-world scenarios.

Despite these advantages, our proposal has certain limitations. First, the blob generation model is trained solely on the BDD dataset, and its capacity on out-of-distribution data is unstable. As illustrated in Fig. 10, we conduct tests on both the KITTI [59] and HDD [60] datasets. Our findings reveal that the generative model struggles to reconstruct objects it has not encountered during training. This limitation is evident in the first two columns, where the generated visualizations exhibit unusual artifacts. However, in more common scenarios, as depicted in the last column, the results are more plausible. This highlights the inherent challenges and potential unreliability of generative models when applied to out-of-distribution samples. Second, during our attribution computation, we accumulate multi-layer information, choosing a combination of one lower and one higher layer, which has proven effective and nearly satisfies the "completeness" axiom. However, there are other layer pairs that could fulfill these criteria. Efficient selection methods other than grid testing remain an area for future research. Another empirical choice is the Kuwahara filter. We find that most edge-preserving filters can enhance the quality of the visualization results, yet identifying an optimal filter efficiently still poses an open question. These limitations highlight potential avenues for future research, which we intend to explore.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] X. Du and K. K. Tan, "Comprehensive and practical vision system for self-driving vehicle lane-level localization," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2075–2088, 2016.

[9] Y. Li, F. Feng, Y. Cai, Z. Li, and M. Á. Sotelo, "Localization for intelligent vehicles in underground car parks based on semantic information," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 1317–1332, 2024.

[10] J. Liu, Y. Cui, J. Duan, Z. Jiang, Z. Pan, K. Xu, and H. Li, "Reinforcement learning-based high-speed path following control for autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 6, pp. 7603–7615, 2024.

[11] J. Wang, H. Sun, and C. Zhu, "Vision-based autonomous driving: A hierarchical reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11 213–11 226, 2023.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] E. W. Saad and D. C. W. II, "Neural network explanation using inversion," *Neural Netw.*, vol. 20, no. 1, pp. 78–93, 2007.

[17] W. Xia, Y. Zhang, Y. Yang, J. Xue, B. Zhou, and M. Yang, "GAN inversion: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3121–3138, 2023.

[18] E. Sotthiwat, L. Zhen, C. Zhang, Z. Li, and R. S. M. Goh, "Generative image reconstruction from gradients," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, 2024.

[19] 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.

[20] H. Yin, A. Mallya, A. Vahdat, J. M. Álvarez, J. Kautz, and P. Molchanov, "See through gradients: Image batch recovery via GradInversion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16 337–16 346.

[21] A. Hatamizadeh, H. Yin, H. Roth, W. Li, J. Kautz, D. Xu, and P. Molchanov, "GradViT: Gradient inversion of vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 011–10 020.

[22] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020.

[23] É. Protas, J. D. Bratti, J. F. de Oliveira Gaya, P. D. Jr., and S. S. C. Botelho, "Visualization methods for image transformation convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 7, pp. 2231–2243, 2019.

[24] 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.

[25] 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.

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

[27] A. M. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3387–3395.

[28] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4217–4228, 2021.

[29] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8107–8116.

[30] A. B. Yildirim, H. Pehlivan, B. B. Bilecen, and A. Dundar, "Diverse inpainting and editing with GAN inversion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 23 063–23 073.

[31] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for StyleGAN image manipulation," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 133:1–133:14, 2021.

[32] Z. Zhang, Y. Yan, J. Xue, and H. Wang, "Spatial-contextual discrepancy information compensation for GAN inversion," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 7432–7440.

[33] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, and M. Cord, "STEEX: Steering counterfactual explanations with semantics," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13672, 2022, pp. 387–403.

[34] 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.

[35] 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.

[36] R. Shi, T. Li, L. Zhang, and Y. Yamaguchi, "Visualization comparison of vision transformers and convolutional neural networks," *IEEE Trans. Multimedia*, vol. 26, pp. 2327–2339, 2024.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.

[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] 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.

[46] 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.

[47] X. Huang, S. Jamonnak, Y. Zhao, T. H. Wu, and W. Xu, "A visual designer of layer-wise relevance propagation models," *Comput. Graph. Forum*, vol. 40, no. 3, pp. 227–238, 2021.

[48] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs," in *Br. Mach. Vis. Conf.*, 2020, pp. 1–13.

[49] G. Papari, N. Petkov, and P. Campisi, "Artistic edge and corner enhancing smoothing," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2449–2462, 2007.

[50] P. Wang and N. Vasconcelos, "A generalized explanation framework for visualization of deep learning model predictions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9265–9283, 2023.

[51] 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.

[52] M. Ancona, C. Öztireli, and M. H. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *Proc. Int. Conf. Mach. Learn.*, vol. 97, 2019, pp. 272–281.

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[54] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[55] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

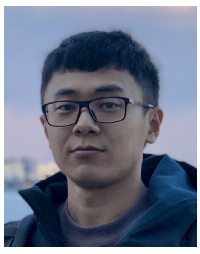
[56] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Representations*, 2018.

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

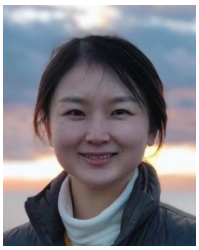
[58] J. Jeon, J. Kim, K. Lee, S. Oh, and J. Ok, "Gradient inversion with generative image prior," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29 898–29 908.

[59] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3354–3361.

[60] V. Ramanishka, Y.-T. Chen, T. Misu, and K. Saenko, "Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7699–7707.



Rui Shi received his Ph.D. degree in graphic and computer sciences from The University of Tokyo, Tokyo, Japan, in 2022. He was a Visiting Researcher with the Department of General Systems Studies, The University of Tokyo. He is currently a Lecturer with the School of Information Science and Technology, Beijing University of Technology, Beijing, China. His current research interests include autonomous driving, computer graphics, 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 with 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 with 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 (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.