IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 March 2024, accepted 15 April 2024, date of publication 19 April 2024, date of current version 29 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391424

RESEARCH ARTICLE

BSED: Baseline Shapley-Based Explainable Detector

MICHIHIRO KUROKI, (Member, IEEE), AND TOSHIHIKO YAMASAKI, (Member, IEEE)
Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

Corresponding author: Michihiro Kuroki (kuroki@cvm.t.u-tokyo.ac.jp)

This work was supported in part by the Institute for AI and Beyond of The University of Tokyo.

ABSTRACT Explainable artificial intelligence (XAI) has witnessed significant advances in the field of object recognition, with saliency maps being used to highlight image features relevant to the predictions of learned models. Although these advances have made artificial intelligence (AI)-based technology more interpretable to humans, several issues have come to light, as some approaches present explanations irrelevant to predictions, and cannot guarantee the validity of XAI (axioms). In this study, we propose the Baseline Shapley-based Explainable Detector (BSED), which extends the Shapley value to object detection for images, thereby enhancing the validity of interpretation. The Shapley value can attribute the prediction of a learned model to a baseline feature while satisfying the explainability axioms. The processing cost for the BSED is within the reasonable range, while the original Shapley value is prohibitively computationally expensive. Furthermore, BSED is a generalizable method that can be applied to various object detectors for images in a model-agnostic manner, and interpret various detection targets without fine-grained parameter tuning. These strengths can enable the practical applicability of XAI. We present quantitative and qualitative evaluations to demonstrate that our method outperforms existing methods in terms of explanation validity. Moreover, we present some applications, such as correcting detection based on explanations from our method.

INDEX TERMS Explainable artificial intelligence, object recognition, Shapley value.

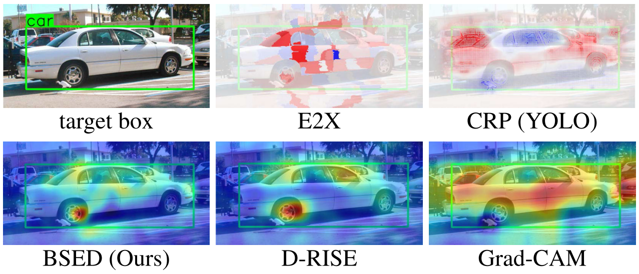
I. INTRODUCTION

Artificial Intelligence (AI)-based object recognition plays an important role across various domains. In the medical field, for instance, AI-based object recognition systems aid physicians in the enhanced and precise diagnosis of diseases. Nonetheless, the black-box nature of AI poses challenges in confidently utilizing such systems. Consequently, the interpretability of AI has drawn prominent academic attention, with extensive research [1], [2] conducted on the topic. This process is critical for the widespread social acceptance of AI systems.

Extensive research has been conducted on explainable artificial intelligence (XAI) for image classification tasks. Typically, pixel-wise feature attributions are calculated based on classification confidence scores and depicted as a saliency map. The feature attribution can be interpreted as an importance score of each pixel for classification. Several different approaches can be employed to calculate the feature attribution. Back-propagation-based methods utilize gradients of the neural network in a learned model, while activation-map-based methods use feature maps in the convolutional neural network layer. Because these methods are dependent upon network architecture, prior knowledge regarding the model is required. In contrast, perturbation-based methods take samples of perturbated input images and their corresponding output scores to calculate the feature attributions. These methods are independent of network architecture, and can be applied in a model-agnostic manner.

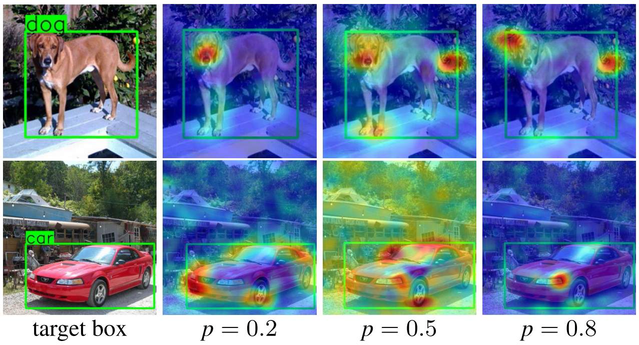
This study focuses on object detection tasks for images, such as that depicted in Fig. 1. The task requires predicting a class label and localizing a bounding box of a target object, thereby making the XAI's task of extracting information from the model more complex. Some methods extend those for image classification by adding conditions on the calculation to limit an explanation scope to a target object.

The associate editor coordinating the review of this manuscript and approving it for publication was Aysegul Ucar.



target box E2X CRP (YOLO)
BSED (Ours) D-RISE Grad-CAM

FIGURE 1. Comparison results with existing methods in interpreting the car detection of YOLOv5s.



target box $p = 0.2$ $p = 0.5$ $p = 0.8$

FIGURE 2. Saliency maps generated from D-RISE explaining detection results. The parameter *p* indicates the percentage of the non-masked area in the images for input samplings.

Contrastive Relevance Propagation (CRP) for the You Only Look Once (YOLO) detector [3] confines the calculation of attributions to those originating from nearby the target bounding box and denoting the class label associated with the target. On the other hand, D-RISE [4] extends the model-agnostic approach [5] and calculates the feature attribution by sampling masked images and their corresponding output scores. The scores take into account both the classification and localization aspects. However, concerns remain about the explanatory validity of these methods. The imposition of the restriction on calculating attributions nearby the target object would introduce bias into the explanations. The result of CRP for the YOLO detector in Fig. 1 shows that other cars within the bounding box have as high feature attributions as the target. The restriction also precludes the possibility of important clues being a little far from the target object. In addition, D-RISE is difficult to generalize to various detection targets, despite a model-agnostic method. Fig. 2 illustrates the changes in saliency maps according to the parameters. A fatal error may appear in an explanation with a certain parameter, and the optimal parameter set may depend on the target, making the method difficult to apply to unknown situations. Therefore, we must pay attention to the explanatory validity. Recently, this topic has frequently been discussed associating with *axioms* [6], [7], which refer to the properties that explainable methods should satisfy. Because the Shapley value [8] has been proven to satisfy the axioms, methods that extend the concept of the Shapley value to XAI have emerged to enhance the validity.

Because research pertaining to the validity of XAI for object detection is scarce, we propose a novel method called the Baseline Shapley-based Explainable Detector (BSED). By extending Baseline Shapley [6] and applying the Shapley value to object detection tasks for images, BSED is expected to yield explanations justified by the axioms. The technical contributions of this study are as follows.

- We developed an XAI for object detection that satisfies the axioms by introducing the Shapley value. We experimentally demonstrated that our method outperforms existing methods in terms of the validity of explanation.
- The inclusion of the Shapley value brings about a significant computational load. To remedy this issue, we introduced the reasonable mathematical Shapley value approximation to reduce the computation burden to a realistic cost while preserving a higher accuracy compared to existing methods.
- We experimentally showed that our method can be applied not only in a model-agnostic manner, but also independently of fine-grained parameter tuning. This indicates that our method can interpret detection results equitably under a wide range of situations.
- We demonstrated the manipulation of detection results according to the explanations of our method, uncovering the potential of extending it to practical applications.

In the remainder of this paper, firstly, we summarize related studies and provide a derivation of our method. Then, we experimentally demonstrate our method's validity and possible applications. Lastly, concluding remarks are presented.

II. RELATED WORKS

A. XAI FOR IMAGE CLASSIFICATION

Back-propagation-based methods [9], [10], [11], [12] utilize gradients of the neural network to calculate pixel-wise feature attribution. For instance, Layer-wise Relevance Propagation (LRP) [13] involves the backward propagation of a classification score (relevance) through the neural network layers, thereby attributing relevance to each pixel in an input image. To distinguish the relevance associated with the true object from other class objects, CRP [14], [15] calculates the disparity between the relevance attributed to the target class and the mean relevances derived from other classes. Integrated Gradients (IG) [7] involves varying the input image from the baseline to the original image and integrating the corresponding gradients. Activation-map-based methods [16], [17], [18], [19] utilize feature maps for the calculation. For instance, Gradient-weighted Class Activation Mapping (Grad-CAM) [20] calculates weighted sums of the feature maps using the gradients within the neural network as the corresponding weights. These methods reduce computational complexity by extracting information from the model. However, they require prior knowledge about the optimal locations for information extraction within the models. In contrast, perturbation-based methods [21], [22], [23] sample perturbed inputs and the model's output for the calculation. For instance, Randomized Input Sampling for Explanation (RISE) [5] samples partially masked images and calculates weighted sums of the input masks using the corresponding output scores as weights. Although they can generate saliency maps without knowledge about the model architecture, they are not considered suitable in cases that require real-time computation, as they sample numerous images.

B. XAI FOR OBJECT DETECTION

Many methods [3], [24] extend XAI for image classification by incorporating conditions for the calculation of feature attributions that specifically target individual objects. CRP for the YOLO [3] calculates the backward propagation of a classification score. This score is associated with a region close to the target bounding box and signifies the target class label. Explain-to-fix (E2X) [25] partitions an image into superpixels and computes the average attributions across them, thereby mitigating pixel-wise noise. Although originally designed for classification, Grad-CAM can be adapted for object detection by identifying the gradients relevant to the task. D-RISE [4] is an extension of a perturbation-based method called RISE [5] and samples binary masks for an input image. These values, 1 or 0, are determined based on probabilities p and $1-p$, respectively. The output score indicates the detection similarity between detections obtained from a masked image and the target detection. A saliency map can be generated by computing the weighted sums of the input masks using the output scores as weights. Fig. 1 presents a comparison of existing methods, with the explanation target being the car detection of a small YOLOv5 (YOLOv5s) [26] model from the Common Objects in COntext (COCO) [27] dataset. The methods presented in Fig. 1 have been curated from diverse categories and serve as baselines for subsequent evaluations. Originally designed for classification tasks, Grad-CAM struggles to target individual objects, instead responding to all objects in the same category. E2X and CRP calculate positive and negative attributions, depicted as red and blue regions. The results of E2X contain a small quantity of noise, and are dependent on the superpixel allocation. Unlike other methods, CRP considers the car's side window as unimportant and other cars within the bounding box as important. D-RISE provides a reasonable result highlighting the important areas for car detection.

C. SHAPLEY VALUE

The Shapley value, originally conceptualized in cooperative game theory [8], has gained prominence in the field of XAI. It offers a systematic method to distribute the "value" or "contribution" of each feature in a prediction model. However, to ensure that the interpretations provided by XAI techniques are meaningful and reliable, they must satisfy certain axioms or principles. A sanity check [28] has highlighted instances where some methods yield explanations that don't align with actual model predictions. Grounded in these concerns, various axioms have been proposed and studied in depth [6], [7], [29]. Here are a few key axioms:

1) DUMMY

If a feature does not influence the output of the score function f, it can be regarded as a dummy and assigned zero attribution.

2) EFFICIENCY

The sum of the attributions across all features should equate to the difference between the output scores of the input $\mathbf{x}$ and the baseline $\mathbf{x^b}$, symbolically:

$$\sum\nolimits_{i=1}^{N} a_i = f(\mathbf{x}) - f(\mathbf{x^b}). \tag{1}$$

This ensures that the entire contribution difference is attributed among the features.

3) LINEARITY

For a linear combination of two score functions, f and g, the attribution of the combined function should equal the sum of the attributions for each individual function. Namely,

$$a_i^{f+g} = a_i^f + a_i^g. \tag{2}$$

Owing to its ability to satisfy these axioms, the Shapley value has been integrated into various XAIs [6], [30], [31], [32]. Despite its advantages, a significant drawback is the computational expense it incurs. While approximations can alleviate this computational demand, they may not always satisfy the axioms, introducing potential discrepancies in the explanations.

III. PROPOSED APPROACH

A. MOTIVATION

From the perspective of practical applicability, existing methods for XAI in object detection have two major issues. The first is the issue of validity, which has been addressed by only a few studies. Fig. 2 shows that the saliency maps generated by D-RISE may vary according to parameters, and we cannot be sure whether a set of parameters is appropriate given an unknown situation. If the method is not sufficiently justified, it cannot be applied in practice. The other primary issue is that of application. Although existing methods can highlight positively important regions, they cannot sufficiently indicate regions of negative importance. In addition, their feature attributions only indicate relative importance, not clarifying how much each area contributes to the prediction. If the explanation is aimed at improving the detector's performance, the degree of impact from positive and negative areas must be important clues for deeper analysis.

These issues have motivated us to develop an *axiom-justified* method that can be generalized to a wide range of situations, and provides more information for deeper analysis.

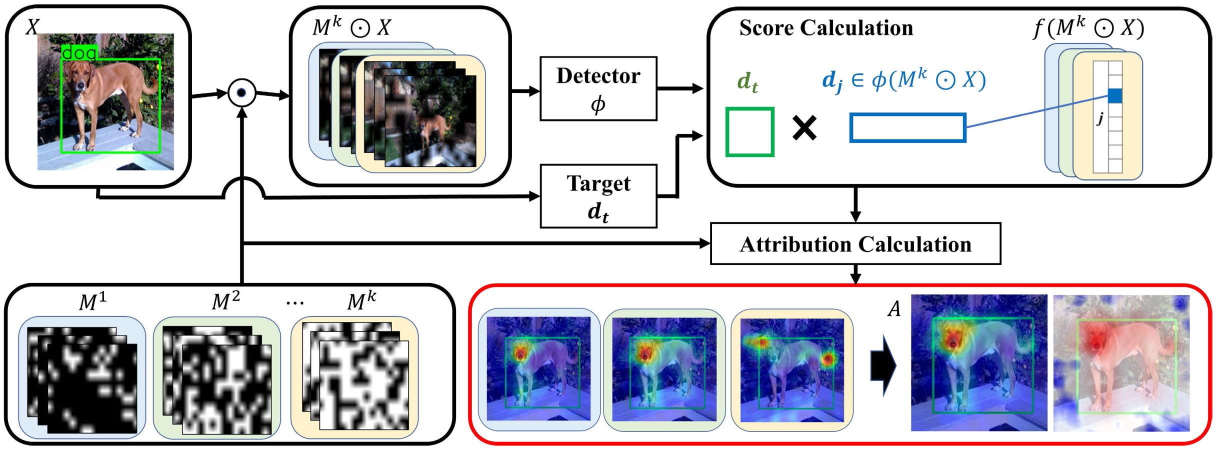


FIGURE 3. Overview of BSED. The explanation target d_t is on the input image X. The detector ϕ obtains perturbated detections d_j from the masked image $M^k \odot X$. The similarity between the explanation target and perturbated detections is obtained to calculate the attribution on each pixel. An attribution map A is generated as an explanation for obtaining the target detection.

In developing this method, we treated the explanation as an attribution task: the entire output score is distributed across each pixel in an image, and the attribution value indicates the contribution of each pixel to the score. Unlike existing methods, we can confirm a balance between positive and negative attributions, employing them as clues to adjust the detection results appropriately.

B. BASELINE SHAPLEY

Let us assume that N_f represents all features of a model's input. Then, the Shapley value can be described as the feature attribution of the target feature $i \in N_f$ as follows:

$$s_i = \sum_{S \subseteq N_f \setminus i} \frac{|S|! * (|N_f| - |S| - 1)!}{|N_f|!} (v(S \cup i) - v(S)). \quad (3)$$

Here, S denotes a subset of features excluding i, and $v(S)$ represents the model output when S is the input. The attribution of i can be obtained by averaging the marginal contributions, which are output changes by the addition of i. In a machine-learning setting, $v(S)$ requires retraining the model using only S as an input, which is extremely time-consuming. SHapley Additive exPlanations (SHAP) [31] avoids this process by approximating $v(S)$ as the expected value of the output scores containing S as an input. However, the expected value significantly depends upon the distribution of the dataset. Therefore, we adopted another Shapley value method that is independent of the distribution of the dataset, namely Baseline Shapley [6], which approximates $v(S)$ by combining the target input $\boldsymbol{x}$ and the baseline input $\boldsymbol{x}^b$ as follows:

$$v(S) = f(\boldsymbol{x}_S; \boldsymbol{x}^b_{N_f \setminus S}). \quad (4)$$

For the input of function f, values corresponding to the feature set S originate from $\boldsymbol{x}$, whereas those corresponding to the other $N_f \setminus S$ originate from the baseline $\boldsymbol{x}^b$. Such an application of Baseline Shapley to object detection has not been attempted previously.

C. BASELINE SHAPLEY FOR OBJECT DETECTION

We take the function f as a score function that includes an object detector, and assume the baseline $\boldsymbol{x}^b$ to be a black image indicating no information. We then can interpret $\boldsymbol{x}_S; \boldsymbol{x}^b_{N_f \setminus S}$ as a masked image, wherein any pixels corresponding to the S are the original pixels, and all other pixels are masked. If we rewrite Eq. 4 with the element-wise multiplication $\odot$, function to generate binary masks $\mathcal{M}(S)$, and image X, we can also reformulate Eq. 3 as follows:

$$a_i = \sum_{S \subseteq N_f \setminus i} \frac{P_r(S)}{|N_f|} \Big\{ f\big(\mathcal{M}(S \cup i) \odot X\big) - f\big(\mathcal{M}(S) \odot X\big) \Big\}, \quad (5)$$

$$P_r(S) = \binom{|N_f| - 1}{|S|}^{-1}. \quad (6)$$

Here, we rewrite s_i as a_i, which is the attribution value corresponding to the feature i. $P_r(S)$ can be represented as the reciprocal of a binomial coefficient. Subsequently, we group S according to the number of non-masked pixels $|S|$.

$$S^l \in \Big\{ S \subseteq N_f \setminus i \;\Big|\; |S| = l \Big\}. \quad (7)$$

We now can transform the summation in Eq. 5.

$$a_i = \frac{1}{|N_f|} \sum_{l=1}^{|N_f|} F_i(S^l), \quad (8)$$

$$\begin{aligned} F_i(S^l) &= \sum_{S^l} P_r(S^l) \{ f(\mathcal{M}(S^l \cup i) \odot X) - f(\mathcal{M}(S^l) \odot X) \} \\ &= \mathbb{E}\big[f(\mathcal{M}(S^l \cup i) \odot X) - f(\mathcal{M}(S^l) \odot X) \big]. \quad (9) \end{aligned}$$

Given that the number of S^l is $\binom{|N_f|-1}{l}$, its reciprocal $P_r(S^l)$ can be considered the event probability of S^l. Therefore, we can approximate $F_i(S^l)$ as the expected value over S^l in Eq. 9. However, we realize that the calculation of Eq. 8 is very time-consuming, requiring $\mathcal{O}(2^{|N_f|})$ inferences to estimate the Shapley value. Therefore, we need reasonable approximations to reduce the calculation cost.

D. APPROXIMATION OF SHAPLEY VALUE

We can interpret Eq. 8 as the average of $F_i(S^l)$ over all the number of pixels. The calculation is redundant because adjacent l would provide similar $F_i(S^l)$. Taking this into consideration, we further reduce the computation by picking up the representative l and approximating the Shapley value by K layers as follows:

$$a_i \approx \frac{1}{K} \sum_{k=1}^{K} F_i(S^k), \tag{10}$$

$$S^k \in \left\{ S \subseteq N_f \setminus i \;\middle|\; \frac{|S|}{|N_f|} = \frac{k}{K+1} \right\}. \tag{11}$$

Here, $F_i(S^k)$ is the expected value of the incremental scores resulting from the participation of i. Inspired by the problem setting of RISE [5], we simplify $F_i(S^k)$ to the expected value of incremental scores between two masks, conditioned on the event that only one of them has an element of 1 on the pixel of i. Here, we rewrite the mask representation from $\mathcal{M}(S^k)$ to M^k, which assigns binaries to all the pixels. If the pixel of i has influential attribution, the contributions of the score change from the participation of i should be highlighted, while those from other pixels are offset. We define the approximated score $\tilde{F}_{i,k} (\approx F_i(S^k))$ using another mask M'^k as follows:

$$\tilde{F}_{i,k} = \mathbb{E}\big[f(M^k \odot X) - f(M'^k \odot X) \;\big|\; M^k(i) - M'^k(i) = 1\big]. \tag{12}$$

We can express the expected value of Eq. 12 as the summation of all combinations of two mask patterns. We describe two binary masks as M_a^k and M_b^k, which have similar patterns to M^k. If they are allowed to be duplicated, we should consider two conditions between the masks, namely $M_a^k(i) - M_b^k(i) = 1$ and $M_b^k(i) - M_a^k(i) = 1$.

$$\begin{aligned}\tilde{F}_{i,k} = \sum_{m_a^k} \sum_{m_b^k} \Big\{ &\big(f(m_a^k \odot X) - f(m_b^k \odot X)\big) \\ &\times P\Big[M_a^k = m_a^k, M_b^k = m_b^k \;\Big|\; M_a^k(i) - M_b^k(i) = 1\Big] \\ &+ \big(f(m_b^k \odot X) - f(m_a^k \odot X)\big) \\ &\times P\Big[M_a^k = m_a^k, M_b^k = m_b^k \;\Big|\; M_b^k(i) - M_a^k(i) = 1\Big]\Big\}.\end{aligned} \tag{13}$$

Here, P indicates the probability. This equation can be further transformed as follows:

$$\begin{aligned}\tilde{F}_{i,k} = \sum_{m_a^k} \sum_{m_b^k} \Big\{ &\big(f(m_a^k \odot X) - f(m_b^k \odot X)\big) \\ &\times \frac{P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_a^k(i) - M_b^k(i) = 1\Big]}{P\Big[M_a^k(i) - M_b^k(i) = 1\Big]} \\ &+ \big(f(m_b^k \odot X) - f(m_a^k \odot X)\big) \\ &\times \frac{P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_b^k(i) - M_a^k(i) = 1\Big]}{P\Big[M_b^k(i) - M_a^k(i) = 1\Big]}\Big\}\end{aligned} \tag{14}$$

$$\begin{aligned}= \sum_{m_a^k} \sum_{m_b^k} \Big\{ &\big(f(m_a^k \odot X) - f(m_b^k \odot X)\big) \\ &\times \frac{P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_a^k(i) - M_b^k(i) = 1\Big]}{P[M_a^k(i) = 1] \cdot P[M_b^k(i) = 0]} \\ &+ \big(f(m_b^k \odot X) - f(m_a^k \odot X)\big) \\ &\times \frac{P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_b^k(i) - M_a^k(i) = 1\Big]}{P[M_b^k(i) = 1] \cdot P[M_a^k(i) = 0]}\Big\}\end{aligned} \tag{15}$$

$$\begin{aligned}= &\frac{1}{P[M^k(i) = 1] \cdot P[M^k(i) = 0]} \\ &\times \sum_{m_a^k} \sum_{m_b^k} \Big\{ \big(f(m_a^k \odot X) - f(m_b^k \odot X)\big) \\ &\times P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_a^k(i) - M_b^k(i) = 1\Big] \\ &+ \big(f(m_b^k \odot X) - f(m_a^k \odot X)\big) \\ &\times P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_b^k(i) - M_a^k(i) = 1\Big]\Big\}.\end{aligned} \tag{16}$$

We can divide the patterns by the combination of $m_a^k(i)$ and $m_b^k(i)$.

$$\begin{aligned}&P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_a^k(i) - M_b^k(i) = 1\Big] \\ &= \begin{cases} P\Big[M_a^k = m_a^k, M_b^k = m_b^k\Big] & \text{if } m_a^k(i) - m_b^k(i) = 1, \\ 0 & \text{if } m_a^k(i) - m_b^k(i) = 0, \\ 0 & \text{if } m_a^k(i) - m_b^k(i) = -1. \end{cases}\end{aligned} \tag{17}$$

$$\begin{aligned}&P\Big[M_a^k = m_a^k, M_b^k = m_b^k, M_b^k(i) - M_a^k(i) = 1\Big] \\ &= \begin{cases} P\Big[M_a^k = m_a^k, M_b^k = m_b^k\Big] & \text{if } m_b^k(i) - m_a^k(i) = 1, \\ 0 & \text{if } m_b^k(i) - m_a^k(i) = 0, \\ 0 & \text{if } m_b^k(i) - m_a^k(i) = -1. \end{cases}\end{aligned} \tag{18}$$

Subsequently, we can reformulate Eq. 16 as follows:

$$\tilde{F}_{i,k} = \frac{G_{i,k}}{P[M^k(i) = 1] \cdot P[M^k(i) = 0]}, \tag{19}$$

$$G_{i,k} = \sum_{m_a^k} \sum_{m_b^k} \Big\{ \Big(f(m_a^k \odot X) - f(m_b^k \odot X) \Big) \times \Big(m_a^k(i) - m_b^k(i) \Big) P\Big[M_a^k = m_a^k, M_b^k = m_b^k \Big] \Big\}. \tag{20}$$

We now seek to reformulate the summation of m_b^k into its expected value.

$$\begin{aligned} G_{i,k} \approx \sum_{m_a^k} \Big\{ & f(m_a^k \odot X) \cdot m_a^k(i) - f(m_a^k \odot X) \cdot \mathbb{E}\big[M_b^k(i)\big] \\ & - \mathbb{E}\big[f(M_b^k \odot X)\big] \cdot m_a^k(i) \\ & + \mathbb{E}\big[f(M_b^k \odot X) \cdot M_b^k(i)\big] \Big\} P\Big[M_a^k = m_a^k\Big] \qquad (21) \\ = \sum_{m_a^k} \Big\{ & \Big(f(m_a^k \odot X) - \mathbb{E}\big[f(M_b^k \odot X)\big] \Big) \\ & \times \Big(m_a^k(i) - \mathbb{E}\big[M_b^k(i)\big] \Big) \Big\} P\Big[M_a^k = m_a^k\Big]. \qquad (22) \end{aligned}$$

In the transformation, we assumed the independence between $f(M_b^k \odot X)$ and $M_b^k(i)$. Given that m_a^k and m_b^k follow the same distribution of M^k, we can rewrite Eq. 22 as follows.

$$\begin{aligned} G_{i,k} \approx & \sum_{m^k} \Big\{ \Big(f(m^k \odot X) - \mathbb{E}\big[f(M^k \odot X)\big] \Big) \\ & \times \Big(m^k(i) - \mathbb{E}\big[M^k(i)\big] \Big) P\Big[M^k = m^k\Big] \Big\} \qquad (23) \\ = & \ \mathbb{E}\big[f\big(M^k \odot X\big) M^k(i)\big] - \mathbb{E}\big[f\big(M^k \odot X\big)\big] \cdot \mathbb{E}\big[M^k(i)\big]. \qquad (24) \end{aligned}$$

By the definition of covariance, we can rewrite the summation as the expected values over M^k. Finally, Eq. 19 is approximated as follows.

$$\tilde{F}_{i,k} \approx \frac{\mathbb{E}\big[f\big(M^k \odot X\big) M^k(i)\big] - \mathbb{E}\big[f\big(M^k \odot X\big)\big] \cdot \mathbb{E}\big[M^k(i)\big]}{\mathbb{E}\big[M^k(i)\big]\Big(1 - \mathbb{E}\big[M^k(i)\big]\Big)}. \tag{25}$$

Because the calculation of exact expected values is difficult, we instead apply a Monte-Carlo approximation.

$$\begin{aligned} \mathbb{E}\big[M^k(i)\big] & \approx \frac{1}{N} \sum_{j=1}^{N} M_j^k(i), \\ & = \overline{M^k(i)}. \end{aligned} \tag{26}$$

Similar approximations are introduced for other expected values. In the approximation, we randomly sample N binary masks. Thus, the final approximated Shapley value is

$$a_i \approx \frac{1}{K} \sum_{k=1}^{K} \frac{\overline{f\big(M^k \odot X\big) M^k(i)} - \overline{f\big(M^k \odot X\big)} \cdot \overline{M^k(i)}}{Z \cdot \overline{M^k(i)} \cdot \big(1 - \overline{M^k(i)}\big)}. \tag{27}$$

Because the changes of a single pixel would have little effect on the output scores, we change pixels per patch in the mask generation. Therefore, the binary masks are initially generated in small grid size, and subsequently expand to the input image size. Thus, the contribution of the score changes should be equally distributed to all the pixels in the patch. Z refers to the number of pixels in the patch and plays the role of normalization factor. The calculation of Eq. 27 can be performed in parallel for all pixels. Consequently, the attribution map A comprising all a_i is expressed as follows:

$$\begin{aligned} A = \frac{1}{K} \sum_{k=1}^{K} & \Big\{ \overline{f\big(M^k \odot X\big) M^k} - \overline{f\big(M^k \odot X\big)} \cdot \overline{M^k} \Big\} \\ & \oslash \Big\{ Z \cdot \overline{M^k} \odot \big(J - \overline{M^k}\big) \Big\}. \end{aligned} \tag{28}$$

Here, J is an all-ones matrix and $\oslash$ is an element-wise division. The number of inferences is reduced from $\mathcal{O}(2^{|N_f|})$ in Eq. 8 to $\mathcal{O}(N)$ in Eq. 28. The overview of the process and the pseudocode of the algorithm are shown in Fig. 3 and Algorithm 1. Because the Shapley value can represent positive and negative attributions, we can achieve the attribution map illustrating positive areas in red and negative areas in blue. Although the definition of the term has not been clearly established, this paper refers to the saliency map as the map representing all values of attributions in the form of a heat map. In general, the introduction of approximations may degrade the accuracy of the Shapley value. We therefore conducted experiments to determine whether the attribution maps maintain high accuracy while still satisfying the axioms.

E. SCORE FUNCTION

The score function f is inspired by the detection similarity of D-RISE [4]. We define the score obtained from a masked image $M^k \odot X$ as follows:

$$f\big(M^k \odot X\big) = \max_{\boldsymbol{d}_j \in \phi(M^k \odot X)} \mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_j), \tag{29}$$

$$\mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_j) = s_{loc}(\boldsymbol{d}_t, \boldsymbol{d}_j) \cdot s_{cls}(\boldsymbol{d}_t, \boldsymbol{d}_j). \tag{30}$$

Here, ϕ denotes the function of the object detector, $\boldsymbol{d}_j$ indicates the vector representation of a detection result, and $\boldsymbol{d}_t$ is the vector representation of the target detection. The similarity between $\boldsymbol{d}_t$ and $\boldsymbol{d}_j$ is denoted as $\mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_j)$. $s_{loc}(\boldsymbol{d}_t, \boldsymbol{d}_j)$ is the Intersection over Union (IoU), which measures the degree of overlap between the areas of $\boldsymbol{d}_t$ and $\boldsymbol{d}_j$. $s_{cls}(\boldsymbol{d}_t, \boldsymbol{d}_j)$ is the classification score of $\boldsymbol{d}_j$ corresponding to the class label of $\boldsymbol{d}_t$. D-RISE [4] employs the cosine similarity of the class probability vectors between $\boldsymbol{d}_t$ and $\boldsymbol{d}_j$ to calculate the similarity of classification, including the probabilities for all classes. However, our method aims to calculate the attribution of the classification score pertaining to the target class. In addition, there are concerns about susceptibility to classes that are unrelated to the target. Therefore, we handle the similarity to the output class itself, rather than that of the probability vectors for all classes.

IV. EVALUATION AND RESULTS

In this section, qualitative and quantitative evaluations are conducted to demonstrate the performance of our method. Before presenting the results, detailed descriptions of the

Algorithm 1 Computing Attribution Map A

```
Require: The number of masks $N$, number of layers $K$, detector function $\phi$, patch size $Z(= c \times c)$, image $X$ with a size of $H \times W$, explanation target detection $\boldsymbol{d}_t$, similarity calculation function $\mathrm{Sim}(\cdot)$, and all-ones matrix $J$.
Ensure: Attribution map $A$
$h \leftarrow \lceil \frac{H}{c} \rceil, w \leftarrow \lceil \frac{W}{c} \rceil, A \leftarrow O$
for $k = 1, \ldots, K$ do
    $p \leftarrow \frac{k}{K+1}, A^k \leftarrow O$
    sum_score $\leftarrow 0$, sum_mask $\leftarrow O$, sum_score_mask $\leftarrow O$
    for $j = 1, \ldots, N$ do
        $M_j^k \leftarrow$ Generate a binary mask with a size of $h \times w$, where the elements are selected as 1 with probability $p$. Subsequently, expand it to the size of $H \times W$ by the bilinear interpolation.
        $f\left(M_j^k \odot X\right) \leftarrow \max_{\boldsymbol{d}_j \in \phi(M_j^k \odot X)} \mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_j)$
        sum_score $\leftarrow$ sum_score $+ f\left(M_j^k \odot X\right)$
        sum_mask $\leftarrow$ sum_mask $+ M_j^k$
        sum_score_mask $\leftarrow$ sum_score_mask $+ f\left(M_j^k \odot X\right)M_j^k$
    end for
    $\overline{f\left(M^k \odot X\right)} \leftarrow$ sum_score$/N$
    $\overline{M^k} \leftarrow$ sum_mask$/N$
    $\overline{f\left(M^k \odot X\right)M^k} \leftarrow$ sum_score_mask$/N$
    $A^k \leftarrow \left\{\overline{f\left(M^k \odot X\right)M^k} - \overline{f\left(M^k \odot X\right)} \cdot \overline{M^k}\right\} \oslash \left\{K \cdot Z \cdot \overline{M^k} \odot \left(J - \overline{M^k}\right)\right\}$
    $A \leftarrow A + A^k$
end for
return $A$
```

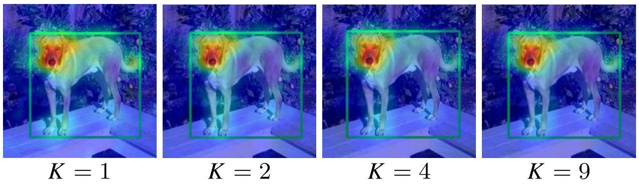


FIGURE 4. Saliency maps targeting the same detection result as shown in Fig. 2, produced by varying the number of layers K. The saliency maps are fairly consistent and appear independent of K to human eyes.

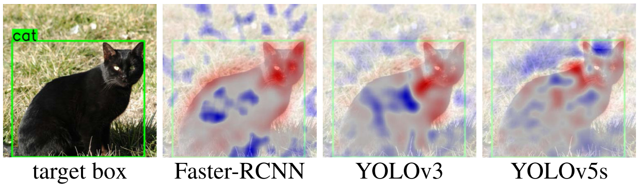


FIGURE 5. Attribution maps on the explanation for obtaining the target detection of a cat. The explanation results from different object detectors are compared.

experimental conditions and evaluation metrics related to these evaluations are provided.

A. PARAMETER CONFIGURATION

In the evaluation of our method, we set the following parameters: patch size $Z = 32 \times 32$, number of masks $N = 6,000$, and number of layers $K = 4$. Unless otherwise noted, all experiments for BSED throughout this study used these parameters. The size of all input images employed in this study were resized to approximately 600 pixels to fully leverage the performance of the object detectors. The patch size Z, which affects the resolution of the attribution maps, depends upon the size of the input image. The values of N and K affect the approximation accuracy of Eq. 28. We consider $N = 6,000$ reasonable, as other explainable methods [4], [5], [33] using Monte Carlo sampling tend to employ parameters of approximately the same magnitude. Note that, the computational cost of BSED increases in proportion to K. The result of Fig. 4 shows that $K = 4$ is sufficiently accurate for human eyes while refraining from the increase in calculation costs.

B. FUNDAMENTAL EVALUATION

1) EVALUATION RESULTS

Fig. 1 shows that BSED effectively captures the car's characteristics, producing less noise compared to other methods. Additionally, BSED is versatile and can be applied to a range of detectors. This includes two-stage detectors, which produce region proposals before classification, and one-stage detectors that manage classification and localization simultaneously. Fig. 5 presents a comparison with the target detection being the ground truth for a cat. The attribution map serves to reflect the detector's performance. Faster-RCNN [34], a representative two-stage detector, can accurately detect the cat, yielding a distinct attribution map. YOLOv5s [26], one of the lightweight one-stage detectors, misclassified the object as a bear, failing to capture the cat's characteristics.

2) COMPUTATIONAL TIME

Our BSED took 174 seconds on a single Nvidia Tesla V100 to generate the saliency map of Fig. 1. In contrast, D-RISE required 47 seconds for the calculation, as it does not employ a multilayer approximation. However, this approximation significantly enhances the accuracy of saliency maps. Users, especially in safety-critical fields, who seek accurate explanations might be willing to allocate more computation time and prioritize reliability over speed. Furthermore, embracing parallel computing and advancements in GPU technology can significantly reduce the computational time. As such, we don't see the processing time of BSED as a major drawback.

3) DEPLOYMENT

Our method is hardware-agnostic, allowing it to be utilized even on an edge device equipped only with a CPU. Regardless of the deployment environment, the computational time is proportional to the inference time of the object detector. Although our method can operate independently on an edge device, it is not intended for real-time use at this stage, and performance analysis offline is the main focus. In cases where high-spec machines with multi-core processors for parallel computation are available, our method can run on them faster by transferring the object detector and image data from the device.

C. EXPERIMENT AND EVALUATION SETTINGS

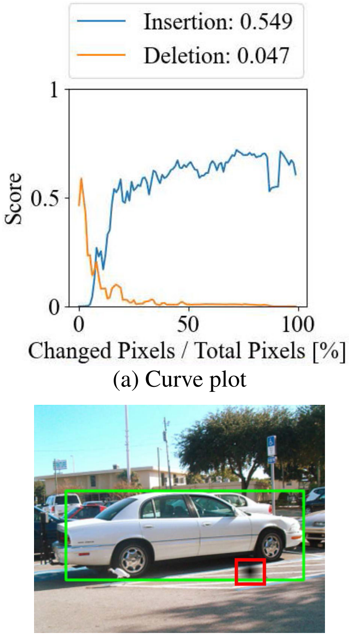
To quantitatively compare our method with the existing methods, we have to select appropriate evaluation metrics. While various evaluation metrics have been proposed, it remains unsettled as to which ones should be consistently employed. Therefore, we selected widely employed metrics in the benchmark evaluation for a fair comparison.

1) ENERGY-BASED POINTING GAME

Pointing Game [35] has been a widely-used evaluation metric in various research, which measures whether the pixel with the highest attribution is within the target bounding box. However, this method does not consider the attribution of other pixels. Therefore, we adopted Energy-based Pointing Game (EPG) [17] to consider the states of other attributions. In this metric, the number of feature attributions gathered in the target bounding box is evaluated using L_{EPG} defined as follows:

$$L_{EPG} = \frac{\sum L_{(i,j)\in bbox}}{\sum L_{(i,j)\in bbox} + \sum L_{(i,j)\notin bbox}}. \tag{31}$$

Here, $L_{(i,j)\in bbox}$ denotes the attribution value of any pixel (i, j) which is located inside the target bounding box. Because some methods compared in the benchmark evaluation calculated negative feature attributions, we normalized the attributions by the min-max normalization for the calculation of Eq. 31.



(a) Curve plot

(b) Example of mask

FIGURE 6. (a) Curve plot for the Deletion and Insertion evaluation derived from the saliency map in Fig. 1. The horizontal axis represents the percentage of the number of removed or added pixels. The AUC of these plots is also shown. (b) Example of a mask applied to the input image in the Dummy evaluation. The red rectangle highlights the location where the mask was applied.

2) DELETION AND INSERTION

We adopted deletion and insertion [38] metrics for the benchmark evaluation because these are also widely adopted by various research. In the deletion metric, pixels with higher attributions are sequentially removed from the input image, and the corresponding decrease in the output score of the model is evaluated. In the insertion metric, pixels are added to the black image in the same order, and the corresponding increase in the output score of the model is evaluated. Fig. 6(a) shows the curve plots of these metrics calculated from the saliency map shown in Fig. 1. The area under the curve (AUC) is an indicator of this quantitative evaluation.

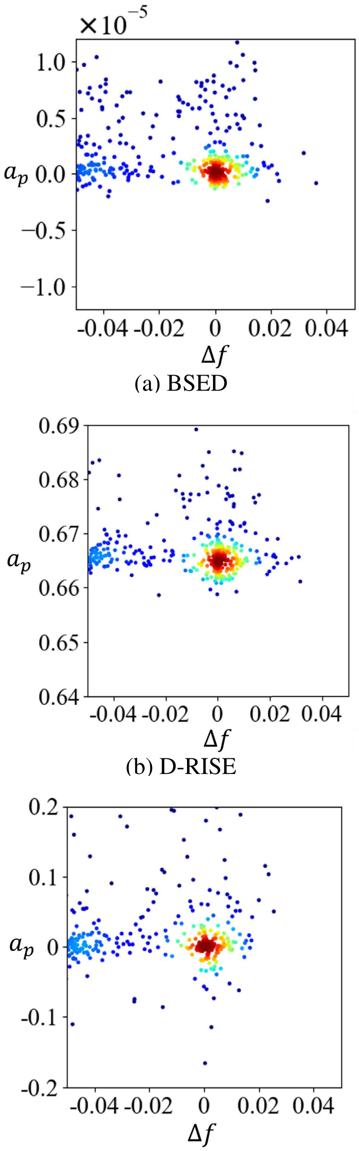
As an evaluation for *axioms* in XAI for object detection has not been established, we conducted new evaluations to confirm whether the method satisfies the axioms.

3) DUMMY

We assessed the assignment of zero attributions for dummy features. When masking a pixel has no impact on the output score, we can interpret that pixel as a dummy feature. To make the score changes observable, we masked pixels in distinct patch regions as shown in Fig. 6(b), evaluated the change in score denoted as Δf, and determined the mean attribution values across the patch, denoted as a_p. The size of the patch is equivalent to the aforementioned Z. Fig. 7 shows the relationship between these values, targeting the saliency map of Fig. 1. The patch regions were randomly created over the entire image. The plots of BSED and E2X are concentrated at the zero value, indicating that most of the dummy pixels

TABLE 1. Results of quantitative evaluation comparison with existing methods. Energy-based Pointing Game, Deletion, and Insertion are denoted as EPG, Del., and Ins. respectively. $\overline{\mathcal{D}(A)}$ and $\overline{\mathcal{E}(A)}$ denote the metrics of *Dummy* and *Efficiency*.

	COCO [27]					VOC [36]				
	EPG(↑)	Del.(↓)	Ins.(↑)	$\overline{\mathcal{D}(A)}$(↓)	$\overline{\mathcal{E}(A)}$(↓)	EPG(↑)	Del.(↓)	Ins.(↑)	$\overline{\mathcal{D}(A)}$(↓)	$\overline{\mathcal{E}(A)}$(↓)
Grad-CAM	0.215	0.142	0.496	0.140	-	0.336	0.128	0.439	0.173	-
E2X	0.131	0.094	0.330	0.016	-	0.239	0.090	0.273	0.019	-
CRP (YOLO)	0.166	0.177	0.461	0.008	-	0.260	0.206	0.359	0.020	-
D-RISE	0.205	0.035	0.636	0.330	-	0.311	0.043	0.551	0.307	-
BSED ($K=1$)	0.211	0.037	0.642	2.49×10^{-6}	0.511	0.314	0.042	0.552	2.51×10^{-6}	0.500
BSED ($K=2$)	0.227	0.034	0.660	1.87×10^{-6}	0.298	0.328	0.041	0.571	1.83×10^{-6}	0.286
BSED ($K=4$)	**0.244**	**0.034**	**0.667**	$\mathbf{1.44\times10^{-6}}$	**0.184**	**0.338**	**0.041**	**0.581**	$\mathbf{1.41\times10^{-6}}$	**0.176**



(a) BSED

(b) D-RISE

(c) E2X

FIGURE 7. Scatter plot showing the relationship between Δf and a_p. The kernel density function was used to indicate high densities in red and low densities in blue.

TABLE 2. Results of quantitative evaluation using the same metrics as in the benchmark evaluation. The explanations of detection results derived from the COCO [27] dataset were evaluated.

	YOLOv3 [37]				
	EPG(↑)	Del.(↓)	Ins.(↑)	$\overline{\mathcal{D}(A)}$(↓)	$\overline{\mathcal{E}(A)}$(↓)
D-RISE	0.219	0.053	0.737	0.475	-
BSED ($K=1$)	0.225	0.055	0.739	2.98×10^{-6}	0.569
BSED ($K=2$)	0.247	0.053	0.762	2.40×10^{-6}	0.324
BSED ($K=4$)	**0.269**	**0.052**	**0.770**	$\mathbf{1.94\times10^{-6}}$	**0.215**
	Faster-RCNN [34]				
	EPG(↑)	Del.(↓)	Ins.(↑)	$\overline{\mathcal{D}(A)}$(↓)	$\overline{\mathcal{E}(A)}$(↓)
D-RISE	0.275	0.213	0.826	0.563	-
BSED ($K=1$)	0.279	0.214	0.826	2.41×10^{-6}	0.573
BSED ($K=2$)	0.304	0.204	0.843	1.84×10^{-6}	0.325
BSED ($K=4$)	**0.327**	**0.201**	**0.853**	$\mathbf{1.43\times10^{-6}}$	**0.182**

have zero attributions. We define the criteria as follows:

$$\mathcal{D}(A) = \overline{|a_d|}, \quad a_d \in \left\{ a_p \;\middle|\; |\Delta f| < \sigma \right\}. \tag{32}$$

Here, A denotes the attribution map, and σ represents a threshold for distinguishing the dummy features. The lower value indicates the method satisfies the dummy property. We calculated the mean $\mathcal{D}(A)$ across all the attribution maps obtained in the benchmark evaluation. We set $\sigma = 0.005$ and denoted the result as $\overline{\mathcal{D}(A)}$ in Table 1, showing that BSED satisfies the dummy property better than other methods.

4) EFFICIENCY

We evaluated whether the sum of the attributions is equivalent to that of the output scores of the input image. We define the efficiency metrics as follows:

$$\mathcal{E}(A) = \left| \sum_{a \in A} a - f(X) \right|. \tag{33}$$

We evaluated the average of $\mathcal{E}(A)$ over all attribution maps obtained in the benchmark evaluation. The result, as $\overline{\mathcal{E}(A)}$ in Table 1, indicates that our method most satisfies the efficiency property among the tested methods. Values for other methods are omitted since they do not meet the efficiency property, resulting in excessive values.

5) LINEARITY

We assessed if the BSED framework satisfies linearity. If we define the similarity as $\mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_j) = s_{loc}(\boldsymbol{d}_t, \boldsymbol{d}_j) + s_{cls}(\boldsymbol{d}_t, \boldsymbol{d}_j)$ in Eq. 30, and denote the most similar detection as $\boldsymbol{d}_{\max}$, the output score $\mathrm{Sim}(\boldsymbol{d}_t, \boldsymbol{d}_{\max})$ can be defined as a linear

TABLE 3. One-sided test for evaluation results conducted in our study.

		P-Value (one-sided)				
Detector	Dataset	EPG	Del.	Ins.	$\overline{\mathcal{D}(A)}$	$\overline{\mathcal{E}(A)}$
YOLOv5s	COCO	<0.001	0.361	<0.001	<0.001	<0.001
YOLOv5s	VOC	<0.001	0.040	<0.001	<0.001	<0.001
YOLOv3	COCO	<0.001	0.172	<0.001	<0.001	<0.001
FasterRCNN	COCO	<0.001	<0.001	<0.001	<0.001	<0.001

TABLE 4. Survey results showing the percentage of each response for all the task images. 54.8% of the total responses indicated our method is much or slightly better than another.

BSED is much better	BSED is slightly better	About the same	D-RISE is slightly better	D-RISE is much better
20.1 %	34.7 %	19.8 %	16.5 %	8.9 %

combination of $s_{loc}(\boldsymbol{d}_t, \boldsymbol{d}_{\max})$ and $s_{cls}(\boldsymbol{d}_t, \boldsymbol{d}_{\max})$. Because the calculation of Eq. 28 includes the linear calculation, the attributions must be linear combinations of those from $s_{loc}(\boldsymbol{d}_t, \boldsymbol{d}_{\max})$ and $s_{cls}(\boldsymbol{d}_t, \boldsymbol{d}_{\max})$.

D. BENCHMARK EVALUATION RESULTS

Quantitative evaluations were conducted using a subset of 10% of images randomly selected from the widely used COCO [27] and PASCAL Visual Object Classes (VOC) [36] datasets of validation splits. The detections from YOLOv5s were set as explanation targets. For a fair comparison, the score functions required for the evaluation metrics and the calculation in D-RISE are identical to that described in Eq. 29. Table 1 indicates BSED exhibited the best performance across all indicators. Additionally, we evaluated BSED using different numbers of layers. As the number of layers increases, the accuracy of the saliency maps likewise increases, demonstrating the efficacy of the multilayer calculation.

1) ADVANTAGE OVER THE EXISTING METHOD

Let us clarify the advantages of our method in comparison to the existing method of D-RISE [4]. It calculates weighted sums of input masks M using their corresponding output scores as weights. D-RISE calculates a saliency map A_D using Monte-Carlo sampling, as follows:

$$A_D = \sum_{j=1}^{N} f\left(M_j \odot X\right) M_j. \tag{34}$$

Here, f, X, and N denote the score function, an input image, and the number of masks in the sampling, respectively. M_j are binary masks generated randomly, with their elements set to 1 with a probability of p. For a fair evaluation, our study employed the same mask generation and score function for D-RISE and BSED. In the official D-RISE study [4], the value of p was set to 0.5, and we used the same value for our evaluation of D-RISE. However, generating masks using a single probability leads to inaccurate saliency maps. Fig. 8 shows examples of the saliency maps generated in the benchmark evaluation. BSED yields interpretable saliency maps, whereas D-RISE fails to generate clear ones. We attribute this difference to the multi-layer approximation of our method. The parameter p determines the proportion of masked regions in the input image in the sampling. Moreover, it significantly affects the distribution of output scores, thereby influencing the appearance of the saliency map. For instance, when the masked regions are exceedingly difficult for the target object, output scores tend to be biased toward lower values, rendering them less informative and leading to generating inaccurate saliency maps. Our method circumvents this error through multiple-layer calculations involving various ratios of masked regions. Fig. 8 shows the box plots of feature attributions calculated in each layer where the masked ratios are different. The layers which significantly affect feature attributions differ per target object, indicating that $p = 0.5$ is unsuitable for certain target objects. These findings clearly show the efficacy of our multilayer calculation.

2) EVALUATION WITH OTHER OBJECT DETECTORS

We conducted a benchmark evaluation using target detections from other object detectors, YOLOv3 [37] and Faster-RCNN [34]. Table 2 indicates that our method maintains the excellent results, as observed in the benchmark evaluation shown in Table 1. We can also confirm the efficacy of the multilayer approximation. These results quantitatively indicate the proposed method can be applied to various detectors in a model-agnostic manner.

3) SIGNIFICANT DIFFERENCE

To confirm significant differences, we conducted a one-sided test between D-RISE and BSED ($K = 4$) for evaluation results in our study. P-values are shown in Table 3. Our BSED is statistically significant ($p < 0.001$) in most cases. For the deletion metrics, removing only a few critical pixels significantly drops the scores and makes the evaluation value, which is the area under the curve shown in Fig. 6(a), close to zero. This makes the differences between the results of the two methods less noticeable.

E. HUMAN-CENTRIC EVALUATION

To clarify the difference between the two methods, we conducted a survey regarding which one is more understandable for humans. Using Amazon Mechanical Turk[1], online users were asked to provide responses regarding the explanation results, which were obtained from D-RISE and BSED in the quantitative evaluation of Table 1. An example of the task images presented to the users is shown in Fig. 9, and the response results are presented in Table 4. For each task image, responses were obtained from 20 users. Among all responses to all the task images, 54.8% indicated that our method

[1] https://www.mturk.com/

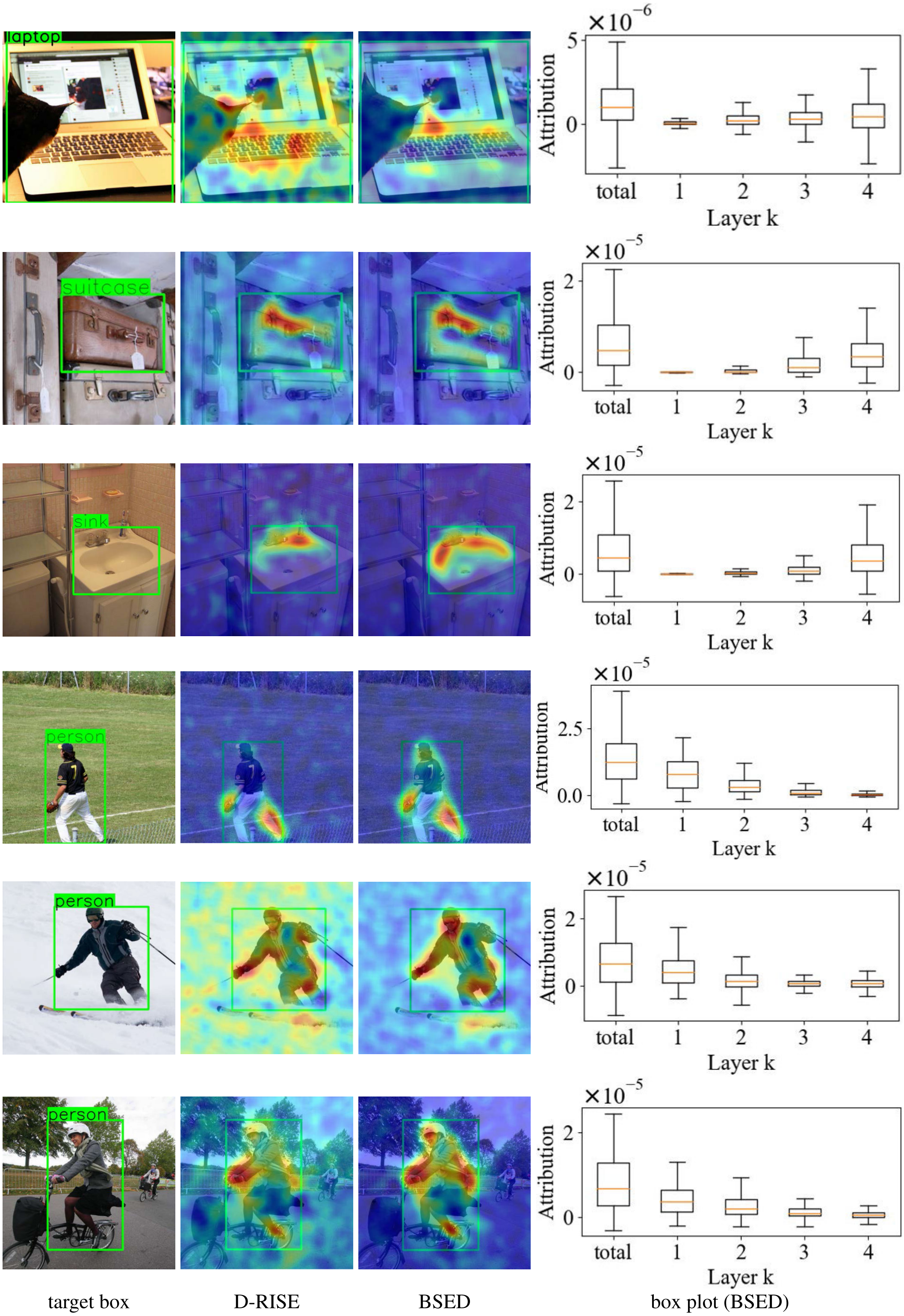
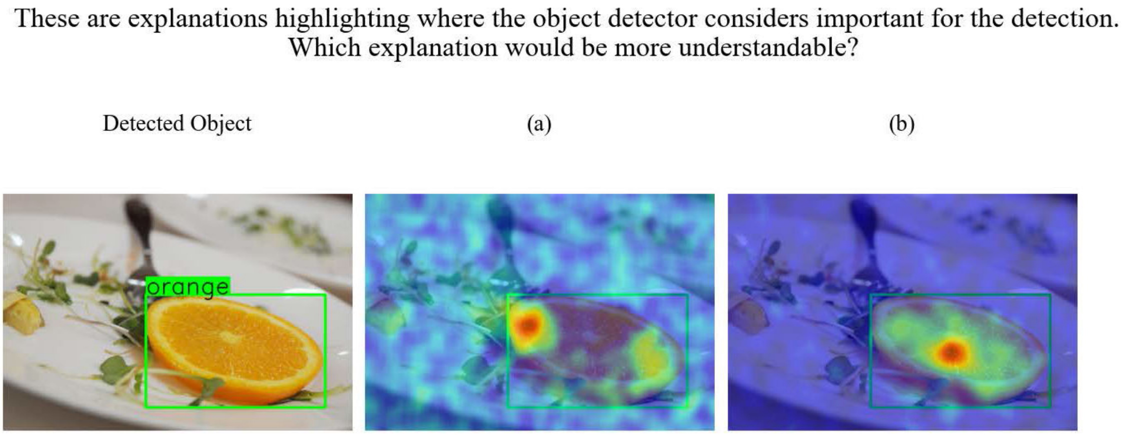


FIGURE 8. Examples of saliency maps obtained in the benchmark evaluation. The rightmost column shows the box plot of feature attributions calculated in each layer by BSED. In the box plot, the orange lines indicate the second interquartile, and the boxes extend from the first to the third interquartile. The topmost and bottommost points indicate the maximum and minimum attribution values, respectively.

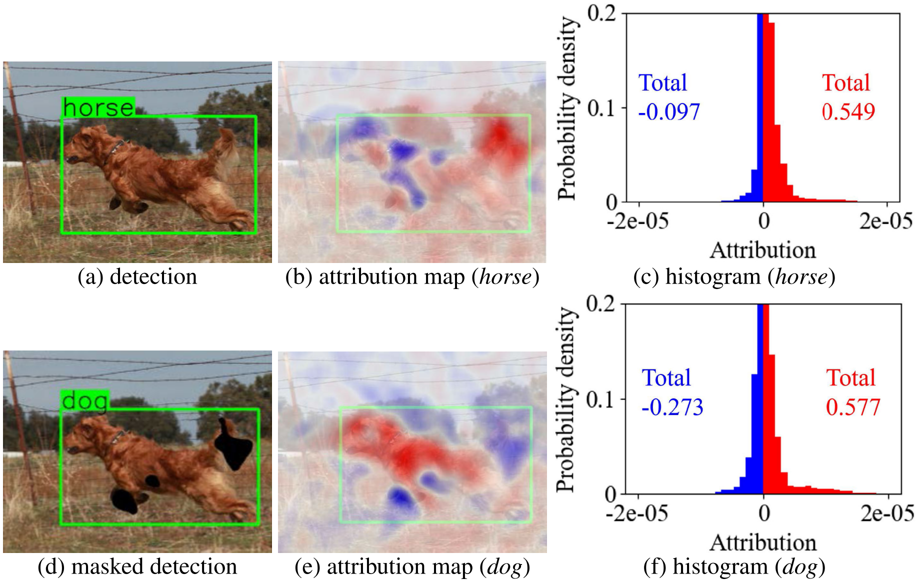
was more understandable, whereas 25.4% favored D-RISE. Additionally, the task images, where responses indicating our method is more understandable outnumbered responses indicating another method is more understandable, accounted



These are explanations highlighting where the object detector considers important for the detection.
Which explanation would be more understandable?

Detected Object (a) (b)

FIGURE 9. Sample of the task images presented to the users during the survey on explanation results. (a) and (b) present explanation results of BSED and D-RISE in random order. The options for response are "(a) is much better", "(a) is slightly better", "About the same", "(b) is slightly better", and "(b) is much better.".



(a) detection (b) attribution map (*horse*) (c) histogram (*horse*)

(d) masked detection (e) attribution map (*dog*) (f) histogram (*dog*)

FIGURE 10. (a) is a misclassified detection. (b) is the corresponding attribution map, and (c) is its histogram. (e) is the attribution map corresponding to a correctly classified detection, and (f) is its histogram. By masking some of the pixels of (e), a corrected detection can be obtained as shown in (d).

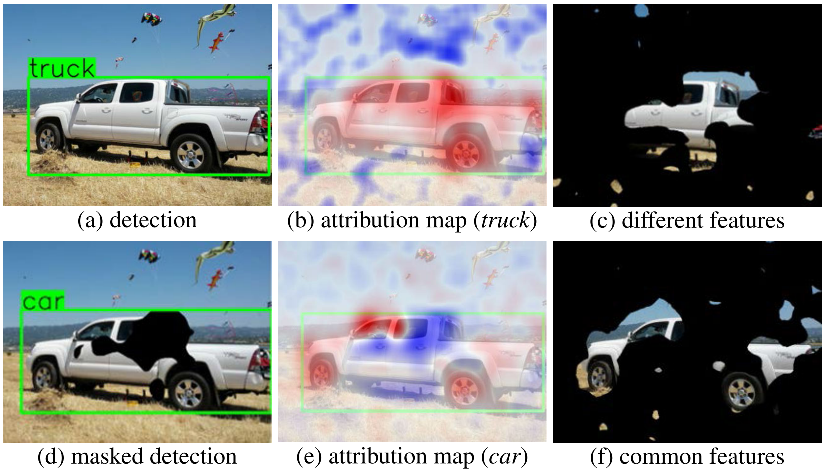
for 74% of the total images. In contrast, the images indicating the opposite trend accounted for 18% of the total.

F. THREAT TO VALIDATION

In the quantitative evaluation, we utilized established evaluation metrics widely used in existing research and newly proposed ones based on the axioms. We consider them appropriate for evaluating the explanatory accuracy. Although our method and D-RISE involve random sampling, they employed a sufficiently large number of samples to ensure convergence of the explanation results. Additionally, combining multiple datasets and object detectors in the evaluation provides multiple perspectives. Our method is model-agnostic, thus applicable to any object detector. We selected representative object detectors from various categories and utilized publicly available implementations. The chosen datasets are widely used for object detection evaluation, containing numerous images depicting various types of objects. Therefore, the quantitative evaluation is both generalizable and reproducible. Given that statistical significance is confirmed in most evaluation results, we consider it valid to conclude that our method outperforms existing methods in terms of explanatory validity.

V. DISCUSSION

In this section, we present detailed analyses of false and true detections as an application of our proposed method.



(a) detection (b) attribution map (*truck*) (c) different features

(d) masked detection (e) attribution map (*car*) (f) common features

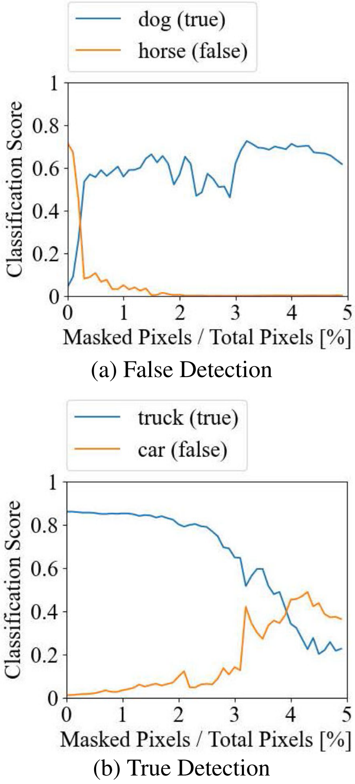
FIGURE 11. (a) is a correct detection, and (b) is the corresponding attribution map. and (e) is the attribution map corresponding to a misclassified detection. (c) depicts the different features between *truck* and *car*, whereas (f) shows the common features between the two.

A. ANALYSIS OF FALSE DETECTION

YOLOv5s mistakenly identified a horse in an image of a dog, as shown in Fig. 10(a). To analyze this misclassification in depth, we generated attribution maps. By assigning the target class labels as *dog* and *horse* in Eq. 30, we derived the corresponding similarity scores. Using these scores, BSED provides explanations for the two corresponding detections, as depicted in Figs. 10(b) and 10(e). The histograms and the sums of the positive and negative attribution values, respectively, are displayed in Figs. 10(c) and 10(f). We identified regions where the detector recognizes features characteristic of dogs and horses. In addition, the histogram of *dog* shows more negative attributions than that of *horse*, implying that the output score of the latter outperforms that of the former. From these results, we infer that the detector does not adequately recognize the features of a dog. The background and lower half of the body appear to contribute to misclassification. By masking pixels exhibiting negative attributions in ascending order, we can increase the classification score of *dog*. Fig. 12(a) shows a plot indicating the increase in the score as pixels with negative values are masked. Only a few pixels must be masked to reverse the score, thereby yielding an accurate detection shown in Fig. 10(d).

B. ANALYSIS OF TRUE DETECTION

YOLOv3 [37] accurately detected a truck, as shown in Fig. 11(a). Given the frequent confusion between trucks and cars, we examine the common features between them. Figs. 11(b) and 11(e) show the reasoning behind the accurate and misclassified detections, respectively. Here, we regard the regions where both attributions for *truck* and *car* are positive as common features. Fig. 11(f) depicts the common features with irrelevant regions masked. These findings, suggesting that the tires and front parts of both *truck* and *car* are similar, are intuitive for human observers. Similarly, regions where only the attribution for *truck* is positive correspond to features exclusive to trucks. This is shown in Fig. 11(c), indicating that the window near the truck's bed is a specific feature for *truck*, which is also interpretable. Finally, an attempt is made to manipulate the true detection and achieve inaccurate classification. Masking the pixels with negative attributions for *car* can increase the score of *car*.



(a) False Detection

(b) True Detection

FIGURE 12. (a) Changes in classification score of the false detection and (b) true detection.

Considering that the different features contribute only to *truck*, masking these features enables the detector to reduce the score of *truck*. Accordingly, masking both feature groups sequentially can effectively increase the score of *car* with a decrease in the score of *truck*. Fig. 12(b) shows a score inversion, resulting in a misclassification with the label *car*, as shown in Fig. 11(d).

VI. CONCLUSION AND FUTURE WORK

Because research on the validity of XAI for object detection is scarce, we proposed BSED, a promising XAI method for object detection that meets the necessary criteria for explainability by incorporating the Shapley value. BSED is not only model-agnostic but also requires no intricate parameter tuning. Through the quantitative evaluation, we demonstrated that our method outperforms the other existing methods in terms of explanatory accuracy across various evaluation metrics. We also showed that by leveraging the attribution maps from our method, detection results can be refined, revealing its potential for real-world applications.

For future work, we aim to continue reducing the computational cost of our method, as it remains high compared to other methods despite its efficient integration of the Shapley value. Balancing both efficiency and explanatory accuracy is crucial for real-world applications. Additionally, we will work on establishing a development process for object detectors using XAI, such as knowledge distillation based on explanation results and retraining focused on challenging scenes. To our best knowledge, BSED is the first XAI for object detection that can be generalizable and precisely quantify both positive and negative prediction contributions. We believe this unique feature will pave the way for innovative advancements in future XAI applications.

ACKNOWLEDGMENT

This work was supported in part by the Institute for AI and Beyond of The University of Tokyo. The subjective evaluation using crowd-sourcing services in this study is approved by the Institutional Review Board (IBR) of the Graduate School of Information Science and Technology, The University of Tokyo: UT-IST-RE-191028-1.

REFERENCES

[1] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019.

[2] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020.

[3] A. Karasmanoglou, M. Antonakakis, and M. Zervakis, "Heatmap-based explanation of YOLOv5 object detection with layer-wise relevance propagation," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Jun. 2022, pp. 1–6.

[4] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. Saenko, "Black-box explanation of object detectors via saliency maps," in *Proc. CVPR*, pp. 11443–11452, 2021.

[5] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proc. BMVC*, 2018, pp. 1–17.

[6] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2020, pp. 9269–9278.

[7] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 3319–3328.

[8] L. S. Shapley, "A value for n-person games," *Contrib. Theory Games*, vol. 2, no. 28, pp. 307–317, 1953.

[9] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Univ. Montreal*, vol. 1341, no. 3, p. 1, 2009.

[10] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," 2017, *arXiv:1706.03825*.

[11] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3145–3153.

[12] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*.

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130140.

[14] J. Gu, Y. Yang, and V. Tresp, "Understanding individual decisions of CNNs via contrastive backpropagation," in *Proc. ACCV*, 2019, pp. 119–134.

[15] H. Tsunakawa, Y. Kameya, H. Lee, Y. Shinya, and N. Mitsumoto, "Contrastive relevance propagation for interpreting predictions by a single-shot object detector," in *Proc. IJCNN*, 2019, pp. 1–9.

[16] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. WACV*, 2018, pp. 839–847.

[17] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-CAM: Score-weighted visual explanations for convolutional neural networks," in *Proc. CVPRW*, 2020, pp. 24–25.

[18] Q. Zheng, Z. Wang, J. Zhou, and J. Lu, "Shap-CAM: Visual explanations for convolutional neural networks based on Shapley value," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 459–474.

[19] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016, pp. 2921–2929.

[20] 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," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[21] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proc. ICCV*, 2019, pp. 2950–2958.

[22] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. SIGKDD*, 2016, pp. 1135–1144.

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 818–833.

[24] T. Yamauchi and M. Ishikawa, "Spatial sensitive GRAD-CAM: Visual explanations for object detection by incorporating spatial sensitivity," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 256–260.

[25] D. Gudovskiy, A. Hodgkinson, T. Yamaguchi, Y. Ishii, and S. Tsukizawa, "Explain to fix: A framework to interpret and correct DNN object detector predictions," 2018, *arXiv:1811.08011*.

[26] Glenn Jocher and Contributors. (2022). *YOLOv5*. [Online]. Available: https://github.com/ultralytics/yolov5

[27] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[28] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. NIPS*, 2018, pp. 9525–9536.

[29] A. Khakzar, P. Khorsandi, R. Nobahari, and N. Navab, "Do explanations explain? Model knows best," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10234–10243.

[30] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 2018, *arXiv:1802.03888*.

[31] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[32] N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath, "FastSHAP: Real-time Shapley value estimation," in *Proc. ICLR*, 2021, pp. 1–23.

[33] D. Schinagl, G. Krispel, H. Possegger, P. M. Roth, and H. Bischof, "OccAM's laser: Occlusion-based attribution maps for 3D object detectors on LiDAR data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1131–1140.

[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NIPS*, 2015, pp. 91–99.

[35] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *Int. J. Comput. Vis.*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018.

[36] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[37] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[38] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3449–3457.

MICHIHIRO KUROKI (Member, IEEE) received the B.E. degree in electrical engineering from Keio University, Japan, in 2013, and the M.E. degree from The University of Tokyo, Japan, in 2015, where he is currently pursuing the Ph.D. degree. Since 2015, he has been an Autonomous Driving Development Engineer with a company. His main research interests include explainable AI for computer vision fields, such as object detection in images and point clouds.

TOSHIHIKO YAMASAKI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from The University of Tokyo. He was a JSPS Fellow for Research Abroad and a Visiting Scientist with Cornell University, from February 2011 to February 2013. He is currently a Professor with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo. His current research interests include attractiveness computing based on multimedia big data analysis and fundamental problems in computer vision and multimedia. He is a member of ACM, AAAI, IEICE, IPSJ, and ITE.