Human Attention-Guided Explainable AI for Object Detection

Guoyang Liu (gyangliu@hku.hk)
Department of Psychology, University of Hong Kong,
Pokfulam Road, Hong Kong

Jindi Zhang (zhangjindi2@huawei.com)
Huawei Research Hong Kong

Antoni B. Chan (abchan@cityu.edu.hk)
Department of Computer Science, City University of Hong Kong,
Kowloon Tong, Hong Kong

Janet H. Hsiao (jhsiao@hku.hk)
Department of Psychology, the State Key Laboratory of Brain and Cognitive Sciences, and the Institute of Data Science,
University of Hong Kong
Pokfulam Road, Hong Kong

Abstract

Although object detection AI plays an important role in many critical systems, corresponding Explainable AI (XAI) methods remain very limited. Here we first developed FullGrad-CAM and FullGrad-CAM++ by extending traditional gradient-based methods to generate object-specific explanations with higher plausibility. Since human attention may reflect features more interpretable to humans, we explored the possibility to use it as guidance to learn how to combine the explanatory information in the detector model to best present as an XAI saliency map that is interpretable (plausible) to humans. Interestingly, we found that human attention maps had higher faithfulness for explaining the detector model than existing saliency-based XAI methods. By using trainable activation functions and smoothing kernels to maximize the XAI saliency map similarity to human attention maps, the generated map had higher faithfulness and plausibility than both existing XAI methods and human attention maps. The learned functions were model-specific, well generalizable to other databases.

Keywords: Object detection; XAI; Human attention; Deep learning; Saliency map

Introduction

In the last decades, deep learning technology has developed tremendously and revolutionized the field of artificial intelligence (AI) (LeCun et al., 2015). Nowadays, deep learning has been widely applied in image classification, object detection, and natural language processing applications (Bashar, 2019). However, the black-box nature and high computational complexity of deep learning models have made their decision-making process opaque to users, significantly affecting user trust and their usefulness. (Rudin, 2019). Although some explainable AI (XAI) methods have been proposed in recent years, they mainly focused on interpreting image classification or natural language processing models. XAI methods for object detection models remained very limited.

A commonly used XAI method for image classification has been to use a saliency map to highlight features contributing to AI systems' decisions. Current methods can be roughly classified into two categories: gradient-based and perturbation-based (Agarwal et al., 2021). Grad-CAM (Selvaraju et al., 2017) and Grad-CAM++ (Chattopadhay et al., 2018) are two representative gradient-based XAI methods for image classification models. However, they are not suitable for object detection models because they can only generate class-specific rather than object-specific saliency maps. RISE (Petsiuk et al., 2018) and D-RISE (Petsiuk et al., 2021) are two widely used perturbation XAI methods, which infer input features contributing to model decisions by perturbing the model input. However, this method does not work well for object detection models because it generates noisy backgrounds and is computationally intensive (Zhao & Chan, 2023; Li et al., 2020). Object detection plays an important role in many critical AI systems, such as autonomous driving (Adarsh et al., 2020; Cai et al., 2021; Chen et al., 2017) and medical diagnosis (Aly et al., 2021; Liu, 2022). Thus, it is essential to develop effective XAI methods for object detection models to make them more useful and accessible to users.

XAI methods are typically evaluated on two main aspects, faithfulness and plausibility. Faithfulness measures how well the highlighted regions of a saliency map reflect features diagnostic to AI's decisions (Chattopadhay et al., 2018; Samek et al., 2016). Faithfulness is typically assessed by examining the amount of change in an AI model's performance when deleting or inserting the highlighted features. Plausibility measures whether the interpretations of AI's operations conform to human cognition (Yin et al., 2022), and it is typically measured by subjective human judgments. Although existing XAI methods can roughly locate the important region for the decision of AI, the low resolution of gradient-based methods and the dispersion intrinsic of perturbation-based methods seriously affect their faithfulness and plausibility (Li et al., 2020). These limitations make current XAI methods hard to be applied to high-level safety-demanded control systems such as autonomous vehicle control (Omeiza et al., 2021). Therefore, we are motivated to design a new XAI method

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

with better faithfulness and plausibility, aiming to address the above limitations.

As for plausibility, recent studies have suggested that the similarity between XAI saliency maps and human attention maps can be used as an objective plausibility measure (Mohseni et al., 2021; Yang et al., 2022). In view of these faithfulness and plausibility issues related to XAI methods for object detection, here we aimed to examine whether we can use human attention during object detection to enhance faithfulness and plausibility of saliency-based XAI methods. Human attention, or more specifically where humans look in a visual task, reflects underlying cognitive processes (e.g., Hsiao, Lan et al., 2021; Chuk et al., 2020). In object detection tasks, or more often referred to as visual search tasks in the cognitive psychology literature, human participants' attention strategies as reflected in eye movement behavior often reflect their sensitivity to features considered relevant to target identification (e.g., Qi et al., 2023a, 2023b; Yang et al., 2023; Hsiao, Chan et al., 2021; see also Hsiao, An et al., 2021; Hsiao & Chan, 2023). Thus, human attention may provide guidance to diagnostic features that are more accessible and interpretable to humans for XAI methods, potentially enhancing both their faithfulness and plausibility.

Accordingly, here we developed the FullGrad-CAM and FullGrad-CAM++ by extending traditional gradient-based methods to generate explanations for object detection models. We further designed a human attention-guided XAI (HAG-XAI) trained with human attention data by using trainable activation functions and smoothing kernels with an objective to maximize the similarity of the generated XAI saliency maps to the human attention maps. Note that here we are learning how to combine the explanatory information in the detector model to best present as an XAI saliency map that is interpretable (plausible) to a human. We then examined whether the resulting saliency map would have enhanced the faithfulness to AI model and the plausibility to humans, and whether the learned weights could be generalized to another object recognition task/image database to enhance faithfulness and plausibility. All learnable parameters in HAG-XAI were interpretable, and thus could help us understand what led to enhanced faithfulness. The remainder of this paper is organized as follows: In Study 1, we evaluated saliency-based XAI methods, compared XAI and human attention maps, and introduced FullGrad-CAM and Full-Grad-CAM++ for object detection. In Study 2, we presented HAG-XAI and assessed its advantages and generalizability.

Study 1: Comparisons between XAI Saliency Maps and Human Attention Maps

Here we focused our examinations on a representative object detection model with a one-stage architecture, Yolo-v5s [1]

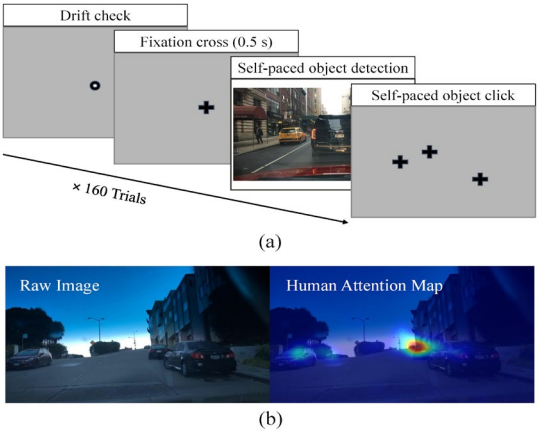

Figure 1: Human attention data collection procedure.

(Jiang et al., 2022; Jocher, 2021). In view of the importance of XAI for object detection during automated driving scenarios to ensure safety (Gupta et al., 2021), we selected the BDD-100K (Yu et al., 2018), a popular well-annotated driving image database, as the target training and evaluation database. All images have a resolution of 1280 x 720. We trained Yolo-v5s using 69,400 images from the training set with five types of labels (including 'car', 'truck, 'bus', 'person', and 'rider') from scratch using the default training configurations. We then tested the trained yolo-v5s model with the validation set (containing 10,000 images), achieving a recall of 75.8%±29.5%. We then randomly selected two independent image subsets (test dataset A and B) from the validation set, each containing 160 images, to conduct the experiments on examining the faithfulness and plausibility of current saliency-based XAI on the model and how they compared with human attention maps.

Methods

Human Attention Data We collected human eye movement data during a vehicle detection task for generating human attention maps. Each trial started with a solid circle at the center of the screen for drift check, followed by a fixation cross for 0.5s. A driving scene image (resized to a resolution of 1024 x 576) was then presented at the center of a 15.6-inch monitor (1920 x 1080-pixel resolution), spanning 34.2° x 20.8° of visual angle under a 55 cm viewing distance. Participants search for vehicle objects (i.e., 'car', 'truck, and 'bus') and remember their locations. They were asked to press the spacebar when they felt they had detected all targets. The screen then turned blank, and participants used a mouse to click on the detected target locations (Figure 1). We recruited 49 participants to perform the task with images in test dataset A, and 27 participants to perform the task with images in test dataset

[1] We also ran experiments on Faster-RCNN and obtained similar results, but do not present them here due to space limitations. More detailed results can refer to Liu et al. (2023).

B. For each subset, eye fixation data of each image over all participants were smoothed by a Gaussian kernel with a standard deviation of 30 pixels, equivalent to one degree of visual angle given the image presentation size.

XAI Methods for Object Detection The vanilla Grad-CAM method (Selvaraju et al., 2017) for image classification AI models highlights regions based on the importance of features with respect to a certain class score. However, it did not consider object detection scenarios. Assuming $M(\cdot)$ is an object detection model such as Yolo-v5s, the output of the model with input image I can be expressed as: $y^m = M(I)$, where y^m with $m = 1, 2\ldots, N_{obj}$ is the output classification probability of m-th detected object, and N_{obj} is the total number of detected objects. If we apply Grad-CAM to an object detection task, the Grad-CAM map for all detected objects can be expressed as:

$$S_G = \sum_{m=1}^{N_{obj}} \mu\left(ReLU\left(\sum_{k=1}^{N_{ch}} \frac{1}{Z} \sum_{ij} \frac{\partial y^m}{\partial A_{ij}^k} A^k \right) \right), \quad (1)$$

where A^k is the activation map in the k-th layer, μ is the max-min normalization function that normalizes the data map to scale between 0 to 1, N_{ch} is the number of channels in A^k, and ReLU is the rectified linear unit activation function. Additionally, Z is a normalization term defined by a global average pooling operation. Chattopadhay et al. (2018) proposed Grad-CAM++ by modifying the gradient term of vanilla Grad-CAM to improve its interpretability. Accordingly, the Grad-CAM++ for object detection can be defined as:

$$S_G^* = \sum_{m=1}^{N_{obj}} \mu\left(ReLU\left(\sum_{k=1}^{N_{ch}} \frac{1}{Z} \sum_{ij} \alpha_{ij}^{km} ReLU\left(\frac{\partial y^m}{\partial A_{ij}^k} \right) A^k \right) \right), \quad (2)$$

where α_{ij}^{km} is a coefficient in (i, j) position for m-th detected object to adjust the weight for k-th channel of the gradient. A ReLU function is applied to the gradient term to retain the most important features with a positive gradient value.

In object detection models, the gradient maps also contain informative spatial information; however, it was not utilized in vanilla Grad-CAM or Grad-CAM++ due to the global average pooling operation applied to gradients (i.e., summation over i and j). As a result, saliency maps generated from Grad-CAM and Grad-CAM++ contain many salient areas not well correlated with detected targets (see Fig. 2). Therefore, here we propose FullGrad-CAM and FullGrad-CAM++ as new XAI methods to generate object-specific saliency maps for object detection models. The FullGrad-CAM method was derived from Grad-CAM, where no average pooling operation is applied to gradients. The FullGrad-CAM is defined as:

$$S_F = \sum_{m=1}^{N_{obj}} \mu\left(ReLU\left(\sum_{k=1}^{N_{ch}} \frac{\partial y^m}{\partial A^k} \odot A^k \right) \right), \quad (3)$$

where $\odot$ represents the Hadamard product.

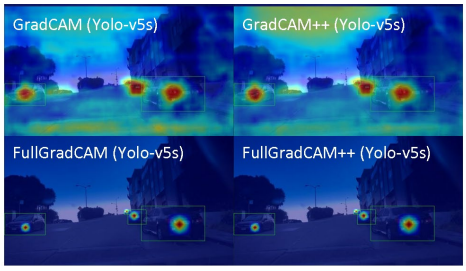

Figure 2: Saliency map examples generated using different XAI methods.

When we apply the ReLU function to the gradient term following Grad-CAM++, our FullGrad-CAM++ is defined as:

$$S_F^* = \sum_{m=1}^{N_{obj}} \mu\left(ReLU\left(\sum_{k=1}^{N_{ch}} ReLU\left(\frac{\partial y^m}{\partial A^k} \right) \odot A^k \right) \right). \quad (4)$$

Previous research (Selvaraju et al., 2017) has suggested that the feature map of the last convolutional layer in deep networks contains the most informative and abstract features. Accordingly, we focused on current examinations on saliency maps generated from the last convolutional layer. For Yolo-v5s, the last convolutional layer of the whole model was used. Note that the last convolutional layer of Yolo-v5s belonged to the neck module, which had a multi-scale branch architecture (i.e., small, middle, and large scales). Hence, we first determined which branch each detected object output was from, and then generated the saliency map accordingly. As shown in Fig. 2, for Yolo-v5s, salient areas were more focused due to small activated areas inside the raw gradient term.[2]

Faithfulness Evaluation Methods We computed the faithfulness using deletion and insertion approaches according to previous studies (Chattopadhay et al., 2018; Petsiuk et al., 2021; Selvaraju et al., 2017). More specifically, the deletion operation deleted salient areas step-by-step according to the saliency scores. The deleted area was filled with random colors. In contrast, the insertion operation inserted salient areas into an empty image with a pure black background step by step according to the saliency scores. For both operations, 100 steps were conducted to record the confidence changes.

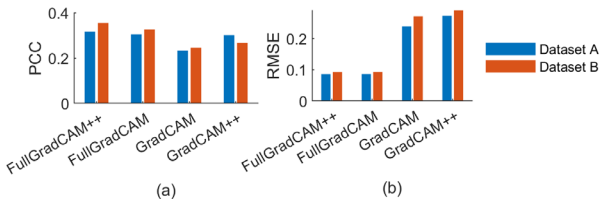

Figure 3: Plausibility of XAI saliency maps for Yolo-v5s using the two test datasets. Higher PCC and lower RMSE indicated better plausibility.

[2] Concurrent work by Zhao & Chan (2023) focuses on instance-specific XAI for object detectors, and uses a similar formulation to (3) but does not sum over the objects. In contrast our work examines the saliency/XAI for all objects in the image, while further comparing to human attention maps for measuring plausibility.

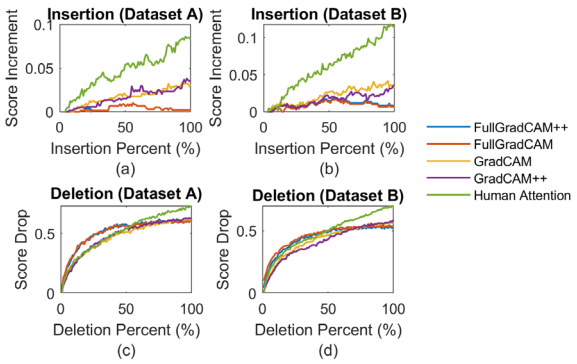

Figure 4: Faithfulness of XAI saliency maps vs. human attention maps for Yolo-v5s using the two test datasets. Higher deletion and insertion scores and larger areas under the deletion/insertion curves indicated higher faithfulness. The Y-axis of the plot denotes the increment or drop of the model prediction score (higher is better).

The maximum deletion and insertion area was limited to the summation area of all detected bounding boxes. In each step, 1% of the total area was deleted or inserted (see Chattopadhay et al., 2018, for details).

Plausibility Evaluation Methods Here we used human attention as an objective human-grounded plausibility criterion. Therefore, the plausibility can be defined as the similarity of the XAI saliency maps to human attention maps. Two similarity measures were employed: (1) Pearson Correlation Coefficient (PCC), which could be seen as a relative similarity measure, and is defined as:

$$PCC = \frac{\mathrm{cov}(u_1, u_2)}{\sigma_{u_1}\sigma_{u_2}}, \tag{5}$$

where u_1 and u_2 are two flattened saliency map vectors, $\mathrm{cov}(\cdot)$ is the covariance function, and σ is the standard deviation function. (2) Root Mean Square Error (RMSE), an absolute similarity measure, is defined as:

$$RMSE = \frac{1}{HW}\left\|u_1 - u_2\right\|_2 \tag{6}$$

where H and W are the height and width of the raw image, and $\|\cdot\|_2$ is the L2-norm operator.

Results

We examined the plausibility and faithfulness of the saliency maps generated using the two proposed FullGrad-CAM methods for the Yolo-v5s model using the two datasets. As measured in PCC and RMSE (Fig. 3), the plausibility of the proposed FullGrad-CAM and FullGrad-CAM++ methods was consistently better than Grad-CAM and Grad-CAM++ methods on the two object detection models and two test datasets. In addition, FullGrad-CAM++ achieved better plausibility than FullGrad-CAM. Regarding faithfulness using the insertion approach (Fig. 4), interestingly, human attention map had higher faithfulness when being used as an XAI saliency map as compared with those from existing saliency-based XAI methods, although our participants had no knowledge of the operations of the AI model. Similarly, using the deletion approach for Yolo-v5s, the final deletion score for human attention is higher than in XAI methods. Note that for smaller deletion percentages, FullGradCAM has higher deletion faithfulness than human attention, which shows that our XAI method does find some specific key features used by the detector.

Together our results suggested that current saliency-based XAI methods for object detection AI models did not capture all the meaningful features used by the models, with their final faithfulness scores lower than attention maps generated by human participants performing the same task but with no knowledge of the models' operations. This result also suggested that the object detection AI model may be using similar information extraction strategies to humans, resulting in high faithfulness when using human attention maps as XAI saliency maps. This finding also justified the use of human attention maps as benchmarks for current saliency-based XAI methods. In addition, the comparison between XAI-generated saliency maps (Fig. 2) and human attention maps (Fig. 1) suggested that smoothing strategies on both gradients and activation may also affect faithfulness. This motivated us in Study 2 to develop an explainable method to find the optimal activation functions and smoothing strategies for enhancing XAI saliency maps' faithfulness and plausibility using human attention data.

Study 2: Human Attention-Guided XAI

This study leverages human attention to optimize gradient-based XAI methods for improving human interpretability. The proposed HAG-XAI incorporates learnable activation functions and smoothing kernels for gradient and activation terms. Using raw activation maps and gradients as input, the model reweights learnable values to enhance plausibility.

In more detail, different from the fixed weight scales for gradient and activation maps used in traditional gradient-based XAI methods, we provide an adaptive piece-wise linear activation function with two learnable parameters:

$$\varphi_{\alpha^-}^{\alpha^+}(\theta) = \alpha^+ \max(\theta, 0) + \alpha^- \min(\theta, 0). \tag{7}$$

The two learnable parameters (α^+ and α^-) allow for different scalings (or complete truncation) of the positive and negative parts of the activation, respectively. For example, the standard ReLU activation is obtained when $\alpha^+ = 1$ and $\alpha^- = 0$. During training, the two parameters used for the activation map (α^+ and α^-) are initialized to 1 (equivalent to a linear activation function), while the two parameters for the gradient map (denoted as β^+ and β^-) are initialized to 1 and 0 (equivalent to ReLU function).

Smoothing kernels are applied to the gradient map and final saliency map to better highlight neighboring features. The gradient can be aggregated over local regions by smoothing to enhance the plausibility. Meanwhile, adding a smoothing operation to the whole saliency map models the difference between the receptive field size of the human and that of the

AI model. The smoothing is implemented with a learnable 2D Gaussian kernel of size 21 x 21,

$$G_A^v(x,y) = A\exp\left(-\frac{(x-x_c)^2+(y-y_c)^2}{2|v|+\varepsilon}\right), \quad (8)$$

where (x,y) is the spatial coordinate, (x_c,y_c) is the constant mean set to the half length of the kernel size (i.e., 11, 11), v is the learnable variance with a initialization value of 3, A is the learnable amplitude with a initialization value of 1, and ε is a small constant to avoid dividing by zero.

Our HAG-XAI saliency generation method is

$$S_{HI} = G_{A_y}^{v_y} * \sum_{m=1}^{N_{obj}} \bar{\mu}\left(ReLU\left(\sum_{k=1}^{N_{ch}}\left(G_{A_\alpha}^{v_\alpha} * \varphi_{\alpha^-}^{\alpha^+}\left(\frac{\partial y^m}{\partial A^k}\right)\right)\odot \varphi_{\beta^-}^{\beta^+}\left(A^k\right)\right)\right), \quad (9)$$

where $\varphi_{\alpha^-}^{\alpha^+}$ and $\varphi_{\beta^-}^{\beta^+}$ are the learnable activations for the gradient and activation map, $G_{A_\alpha}^{v_\alpha}$ and $G_{A_y}^{v_y}$ are learnable Gaussian smooth kernels for the final map and the gradient map, and * is the convolution operator. Note that the same kernel is applied to each channel of gradient and activation tensors. Meanwhile, $\bar{\mu}$ is a normalization function. During visual search, human participants tended to attend to small objects more than large objects. Accordingly, we normalize each object's individual saliency map using their activated area:

$$\bar{\mu}(\theta) = \frac{\theta}{\sum_{ij}\theta+\varepsilon}, \quad (10)$$

where ε is a small constant value to avoid the denominator being 0. A total of 8 learnable parameters are optimized.

The training goal of this model is to obtain a human-like saliency map. Therefore, the loss function is set to the (dis)similarity between human attention map and AI saliency map, based on PCC and RMSE. The optimization objective of the model is

$$\arg\min_{\theta}\left\{1-\frac{\mathrm{cov}\left(S_{HI}^*, S_H^*\right)}{\sigma_{S_{HI}^*}\sigma_{S_H^*}}+\frac{1}{HW}\left(\left\|S_{HI}^*-S_H^*\right\|_2\right)^2\right\}, \quad (11)$$

where S_{HI}^* and S_H^* are the flattened XAI saliency map and human attention map (serving as ground-truth) in the training set.

To test our method, we first used the BDD-100K database used in Study 1 to train the HAG-XAI model and test its faithfulness and plausibility. During training, test dataset A from Study 1 was used as the training/validation set and dataset B as the testing set. The training set is divided into five parts to conduct a five-fold cross-validation for learning the HAG-XAI parameters. All results were from the testing set if not specified. Considering the neck module of Yolo-v5s has three different scales, the activations and gradients were resized to a uniform (maximum) spatial resolution before training.

The Adam optimizer was used during training. The mini-batch size was set to 30. The learning rate was set to 0.05 initially and exponentially decreased to 0.005 within 120 epochs. To avoid overfitting, an early stop strategy was employed, where the patience was set to 30 epochs.

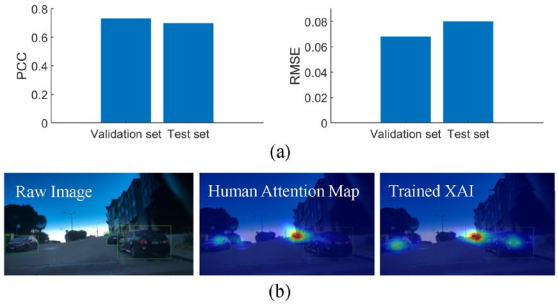

Figure 5: (a) Similarity between the trained HAG-XAI saliency maps and human attention maps using different datasets. (b) Example HAG-XAI saliency maps vs. human attention maps.

Finally, we evaluated the generalization ability of the learned parameters in HAG-XAI, using the validation set of the MS-COCO object detection database, which contained 5000 images with 80 general object classes (Lin et al., 2014).

Results

After training, the well-trained models were assessed on both validation set of dataset A and testing set (dataset B). On dataset A, the averaged 5-fold cross-validation accuracy was reported. The five models generated from the 5-fold cross-validation procedure were assessed on the whole testing set, and the averaged testing accuracy was reported. As shown in Fig. 5a, the performance difference between the validation and testing sets was small, demonstrating good generalization ability. Also, the saliency maps generated from HAI-XAI and human attention maps had high similarity (above 0.7 in PCC on the testing set; Fig. 5b).

The similarity between XAI saliency maps and human attention maps across different XAI methods is depicted in Fig. 6. Saliency maps from HAG-XAI had the highest similarity, indicating a high plausibility. The HAG-XAI also had higher faithfulness than other XAI methods (Fig. 7). These results suggest that human attention maps could be used to guide the design of saliency-based XAI to enhance its faithfulness and plausibility for object detection. Compared with the untrained saliency map shown in Fig. 2, the trained saliency map shown in Fig. 5b was more similar to the human attention maps.

We then used images from MS-COCO database to examine whether the learned functions from HAG-XAI could be transferred to other object detection tasks of the same models.

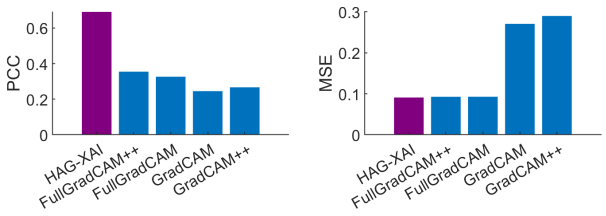

Figure 6: The similarity between XAI saliency maps and human attention maps across different XAI methods.

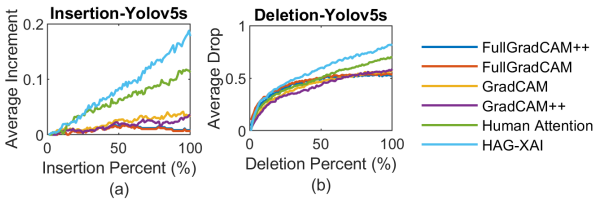

Figure 7: The faithfulness performance comparison between different XAI methods.

Since the image resolutions in MS-COCO were different from BDD-100K, the resolutions of the activation and gradient maps from MS-COCO were adjusted to be equivalent size of BDD-100K. Table 1 illustrates the area under the insertion curve (i-AUC) and deletion curve (d-AUC) for assessing faithfulness of the generated saliency maps over 5000 images with all categories in the database. The results showed that our method achieved the best i-AUC and d-AUC scores in object detection tasks, demonstrating a great generalization ability of HAG-XAI to other databases/object detection tasks.

Discussion

This paper aims to design a new XAI method with high faithfulness and plausibility that can be used in object detection scenarios. To better understand how the learned functions enhance the faithfulness and plausibility of XAI saliency maps, we visualized the learnable parameters since they are fully interpretable. As shown in Fig. 8a, a large smoothing kernel was needed for the gradient term. This is because we used the last convolutional layer of the whole model to generate the saliency map, and the salient area was relatively small in the raw gradient term according to the backpropagation algorithm. In contrast, the global smoothing kernel seemed unnecessary, suggesting that the activation map already matched human attention strategies well. As shown in Fig. 8b, the negative parts of the activations and gradients were turned positive, suggesting that these negative parts also play an important role in explanation, e.g., as counterfactual information. Note that for Yolo-v5s, the outputs of the last convolutional layer used a leaky-ReLU activation function with a leaky factor of 0.1. Therefore, there existed negative values in the activations of Yolo-v5s. Together these visualization results suggested that HAG-XAI provided model-specific guidance for generating XAI saliency maps with high plausibility and faithfulness, with the learning functions well generalizable to other detection tasks using different databases.

Table 1. Faithfulness score when using the learned functions from HAG-XAI (trained on BDD-100K) to generate saliencies for object detection in MS-COCO.

Faithfulness	FGC*	FGC	GC*	GC	Our
d-AUC	0.5206	0.5206	0.7292	0.7413	**0.756**
i-AUC	0.0040	0.0040	0.1032	0.1092	**0.133**

Note: FGC*: FullGrad-CAM++. FGC: FullGrad-CAM.

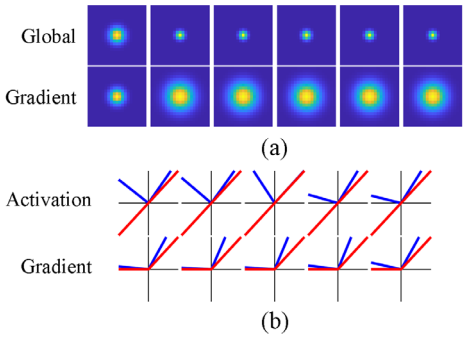

Figure 8: Visualization of HAG-XAI parameters. The left column of (a) is the initial (untrained) Gaussian kernels with a variance of 3, and the other five columns correspond to the learned kernels from five-fold models. The red lines in (b) are the initial (untrained) activation functions, while the blue lines are the trained activations.

In HAG-XAI, three functions—a Gaussian smoothing function, a learnable activation function, and an area-based normalization function—were designed to improve faithfulness and plausibility. To examine the effectiveness of these functions, ablation studies were conducted. As shown in Table 2, the XAI method with all three functions achieved the best faithfulness and plausibility. The Gaussian smoothing function contributed the most; without the function, the faithfulness and plausibility significantly dropped.

In this work, we proposed two novel XAI methods that can generate explanations for object detection models and showed the potential of human attention maps in enhancing the faithfulness of XAI methods. Using human attention maps as guidance, we designed a HAG-XAI method, achieving higher faithfulness and plausibility for object detection models than the existing methods. Meanwhile, the HAG-XAI has the potential to be used as a human attention imitator for object detection tasks (Yang et al., 2022). In future work, we will combine multiple feature maps, rather than only use the last feature map, to extract fine-grained attention of the object detection model, aiming to further enhance the performance of HAG-XAI and explore the potential of the HAG-XAI to be used as a human attention imitator for object detection.

Table 2 Ablation study for Yolo-v5s

Function			Faithfulness		Plausibility	
G	***φ***	***μ***	**d-AUC**	**i-AUC**	**PCC**	**RMSE**
×	×	×	0.4482	0.0099	0.3550	0.0914
√	×	×	0.5612	0.0805	0.6683	0.0900
×	√	×	0.3907	0.0019	0.2947	0.0986
×	×	√	0.3916	0.0021	0.3140	0.0944
√	√	×	0.5474	0.0793	0.6665	0.0871
×	√	√	0.5330	0.0000	-0.3150	0.9643
√	×	√	0.5618	0.0822	0.6873	0.0812
√	√	√	**0.5662**	**0.0843**	**0.6910**	**0.0800**

Acknowledgments

We are grateful to Huawei and RGC of Hong Kong (Collaborative Research Fund No. C7129-20G to Dr. J. Hsiao). We thank Yumeng Yang and Yueyuan Zheng for their help in data collection.

References

Adarsh, P., Rathi, P., & Kumar, M. (2020). YOLO v3-Tiny: Object Detection and Recognition using one stage improved model. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS),

Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S., & Lakkaraju, H. (2021). Towards the unification and robustness of perturbation and gradient based explanations. International Conference on Machine Learning,

Aly, G. H., Marey, M., El-Sayed, S. A., & Tolba, M. F. (2021). YOLO based breast masses detection and classification in full-field digital mammograms. Computer Methods and Programs in Biomedicine, 200, 105823.

Bashar, A. (2019). Survey on evolving deep learning neural network architectures. Journal of Artificial Intelligence, 1(02), 73-82.

Cai, Y., Luan, T., Gao, H., Wang, H., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2021). YOLOv4-5D: An effective and efficient object detector for autonomous driving. IEEE Transactions on Instrumentation and Measurement, 70, 1-13.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE winter conference on applications of computer vision (WACV).

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Chuk, T., Chan, A. B., Shimojo, S., & Hsiao, J. H. (2020). Eye movement analysis with switching hidden Markov models. *Behavior Research Methods*, *52*(3), 1026-1043.

Gupta, A., Anpalagan, A., Guan, L., & Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array, 10, 100057.

Hsiao, J. H., An, J., Zheng, Y., & Chan, A. B. (2021). Do portrait artists have enhanced face processing abilities? Evidence from hidden Markov modeling of eye movements. *Cognition*, *211*, 104616.

Hsiao, J. H., & Chan, A. B. (2023). Visual attention to own- vs. other-race faces: Perspectives from learning mechanisms and task demands. *British Journal of Psychology*.

Hsiao, J. H., Chan, A. B., An, J., Yeh, S.-L., & Jingling, L. (2021). Understanding the collinear masking effect in visual search through eye tracking. *Psychonomic Bulletin & Review*, *28*, 1933-1943.

Hsiao, J. H, Lan, H., Zheng, Y., & Chan, A. B. (2021). Eye Movement analysis with Hidden Markov Models (EMHMM) with co-clustering. *Behavior Research Methods*, *53*, 2473-2486.

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022). A Review of Yolo algorithm developments. Procedia Computer Science, 199, 1066-1073.

Jocher, G. (2021). Yolo-v5. https://github.com/ultralytics/yolov5

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Li, X.-H., Shi, Y., Li, H., Bai, W., Song, Y., Cao, C. C., & Chen, L. (2020). Quantitative evaluations on saliency methods: An experimental study. arXiv preprint arXiv:2012.15616.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. European conference on computer vision,

Liu, G., Zhang, J., Chan, A. B., & Hsiao, J. H. (2023). Human Attention-Guided Explainable Artificial Intelligence for Computer Vision Models. arXiv preprint arXiv: 2305.03601.

Liu, K. (2022). STBi-YOLO: A Real-Time Object Detection Method for Lung Nodule Recognition. IEEE Access, 10, 75385-75394.

Mohseni, S., Block, J. E., & Ragan, E. (2021). Quantitative evaluation of machine learning explanations: A human-grounded benchmark. 26th International Conference on Intelligent User Interfaces,

Omeiza, D., Webb, H., Jirotka, M., & Kunze, L. (2021). Explanations in autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(8), 10142-10162.

Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. arXiv preprint arXiv:1806.07421.

Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., & Saenko, K. (2021). Black-box explanation of object detectors via saliency maps. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Qi, R., Zheng, Y., Yang, Y., Cao, C. C., & Hsiao, J. H. (2023a). Explanation Strategies for Image Classification in Humans vs. Current Explainable AI. arXiv preprint arXiv:2304.04448.

Qi, R., Zheng, Y., Yang, Y., Zhang, J., & Hsiao, J. H. (2023b). Individual differences in explanation strategies for image classification and implications for explainable AI. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems, 28(11), 2660-2673.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision,

Yang, A., Liu, G., Chen, Y., Qi, R., Zhang, J., & Hsiao, J. H. (2023). Humans vs. AI in Detecting Vehicles and Humans in Driving Scenarios. *Proceedings of the 45th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Yang, Y., Zheng, Y., Deng, D., Zhang, J., Huang, Y., Yang, Y., Hsiao, J. H., & Cao, C. C. (2022). HSI: Human Saliency Imitator for Benchmarking Saliency-Based Model Explanations. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing,

Yin, F., Shi, Z., Hsieh, C.-J., & Chang, K.-W. (2022). On the Sensitivity and Stability of Model Interpretations in NLP. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),

Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., & Darrell, T. (2018). Bdd100k: A diverse driving video database with scalable annotation tooling. arXiv preprint arXiv:1805.04687, 2(5), 6.

Zhao, C., & Chan, A. B. (2023). ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection. The Eleventh International Conference on Learning Representations,