Explaining Object Detection Through Difference Map

Shujun Xia[1] Chenyang Zhao[1] Antoni Chan[1]
[1]Department of Computer Science, City University of Hong Kong
{shujunxia2-c, chenyangzhao2-c}@my.cityu.edu.hk, abchan@cityu.edu.hk

Abstract

Visual explanation methods have been effective in interpreting the outputs of object detectors by highlighting important regions corresponding to each model prediction. However, existing approaches have largely overlooked inter-object relationships—particularly the relative importance of each pixel across different objects, a concept we refer to as Object Discrimination (OD). In this paper, we propose difference maps, a novel visual explanation technique designed to enhance the interpretability of object detectors with respect to OD. Serving as a complementary tool to existing instance-specific heat maps, difference maps improve their ability to isolate the impact of key features of individual objects on model outputs. Our qualitative and quantitative evaluation results show that the proposed difference maps can effectively distinguish key features specific to the target object, capturing the relative importance of each pixel across different predictions within the same scene. Our method is applicable to a wide range of object detectors, including one-stage, two-stage, and transformer-based architectures. Furthermore, it enhances existing heatmap-based visual explanation methods by improving focus on the detected object. These results demonstrate the utility of our approach in improving model transparency and interpretability across different detection architectures and explanation techniques.

1. Introduction

Deep neural networks (DNNs) [10] have driven significant progress in object detection and numerous computer vision tasks, demonstrating remarkable performance across a variety of applications. Despite these advancements, understanding the decision-making process of DNNs remains challenging due to their complex and often opaque structures. As spatial convolution is integral to many leading models in computer vision, this has inspired the development of class-specific attention mechanisms [1, 27, 40] to enhance the interpretability of convolutional neural networks (CNNs). These methods generate visual heat maps that highlight important regions in input images, thereby elucidating the contributions of individual pixels to model predictions.

Building on the efforts to interpret CNNs, researchers have developed various approaches to generate visual explanation for deep learning models including object detectors. Perturbation-based methods [3, 5, 8, 15, 20, 23, 26, 32] systematically modify parts of an input image to identify regions that significantly impact model decisions. Class activation map-based (CAM-based) methods [4, 6, 27, 33, 37, 40] adopt weights that reflect the importance of each feature in the feature activation maps to produce heat maps. Furthermore, gradient-based methods [4, 27, 28, 38] analyze the gradients of detector outputs to reveal critical features influencing predictions. For object detectors, instance-specific explanations effectively highlight important regions for each prediction [23, 36, 37, 37, 38]. However, these methods typically neglect the relative importance of each pixel across different objects—meaning that the highlighted regions may be simultaneously important for multiple objects within the same scene. In scenarios where certain pixels are more critical for one object than for others, this reflects a specific form of inter-object relationship, which we term Object Discrimination (OD). Such object relationships can play a critical role in detection outcomes, as interactions between objects can influence how models interpret individual object features and make predictions [14]. Nevertheless, the key limitation of current explanation methods lies in their insufficient exploration and explanation of inter-object relationships.

To address this gap, our work focuses specifically on OD, aiming to identify which object was actually detected by distinguishing it from other objects within the scene. To visualize this relationship, we propose the difference map, which highlights regions that are specifically important for detecting the target object as opposed to its neighbors. This approach enhances the interpretability of object detectors by incorporating object-specific discriminative factors into the explanation. The main contributions of this work are two-fold:

1. We enhance the explainability of object detectors by

considering Object Discrimination, and propose the difference map to isolate the influence of the key features for the prediction of a specific object instance from other objects within the same scene.
2. We demonstrate the effectiveness and generality of our proposed method through both qualitative and quantitative evaluations. Our approach is applied across various types of object detectors, including one-stage, two-stage, and transformer-based architectures. In addition, we integrate the difference map with multiple heatmap-based visual explanation methods and show that it improves the original heatmaps by enhancing focus and reducing irrelevant highlights.

2. Related Works

Object detection. Object detectors typically consist of three main components: a backbone, a neck, and a head. Depending on the type of head used, detectors are generally categorized into one-stage and two-stage methods. Two-stage detectors generate region proposals first, followed by utilizing Region of Interest (RoI) features for subsequent object classification and localization refinement. Prominent examples of this category include R-CNN series, such as R-CNN [10], Fast R-CNN [9], Faster R-CNN [25], and Mask R-CNN [13]. In contrast, one-stage detectors eliminate the need for RoI feature extraction, directly performing object classification and localization on the entire feature map. Representative approaches include YOLO [24], RetinaNet [18], and FCOS [29]. More recently, transformers have been successfully integrated into object detection architectures. They have been used both as a backbone for feature extraction, exemplified by the Pyramid Vision Transformer (PVT) [35], and as detector heads, such as in DETR (DEtection TRansformer) [2]. Our methods offer explanations for both one-stage and two-stage detectors, as well as transformer-based detectors.

Visual explanation for object detection. Visual explanation serves as an effective and intuitive method for interpreting deep learning models, including object detectors. A common approach to generating such explanations is through heat maps, which highlight the important pixels that led to the model's prediction. Perturbation-based methods like D-RISE [23], CAM-based approaches like Spatial Sensitive Grad-CAM (SSGrad-CAM) [37], and gradient-based methods such as Object Detector Activation Map (ODAM) [38] are notable examples that produce instance-specific heat maps for object detectors. D-RISE [23] is a black-box perturbation-based approach, but is computationally demanding and can be prone to noise. SSGrad-CAM [37] and SSGrad-CAM++[36] build upon Grad-CAM[27] and Grad-CAM++ [4], respectively, by incorporating spatial maps to generate more refined, instance-specific heat maps. ODAM [38] leverages gradients of detector targets with respect to feature maps to illustrate how different regions influence the detector's decision for each predicted attribute. In this paper, our proposed method mainly builds upon gradient-based ODAM [38], which was designed to explain a single object instance, to explain object-object interactions. Specifically, we enhance the visual interpretability of object detectors by isolating key features that influence one object over other objects. Our method is general and we show it can also be applied to other heatmap-based visual explanation methods.

Counterfactual explanation. Counterfactual explanations (CEs) aim to clarify why an object is classified as a target class rather than an alternative class, typically by identifying discriminative features [34]. For example, some CEs generate image transformations that shift classification from one class to another, including adversarial attacks or generative approaches that seek realistic but perturbed images [7, 11, 19, 30, 31]. However, many generative methods are limited to simple datasets due to the difficulty of synthesizing realistic images, and exhaustive feature searches are often too computationally expensive for real-time use [11, 19, 21]. In contrast, our paper aims to explain the relative importance of pixels to different objects in the same image during detection. While CEs discriminate between classes, our difference map highlights features that distinguish one object from others within the scene. Importantly, our method is not counterfactual, as it does not rely on alternative class labels. Exploring the integration of our approach with counterfactual explanations presents a promising direction for future research.

3. Methodology

We propose difference maps for interpreting Object Discrimination (OD), which is introduced to highlight pixels that uniquely influence the detection output of a specific object, isolating its key discriminative features from other detected objects. This approach allows for a refined understanding of individual object contributions in complex scenes. Difference maps can be generally applied to a wide range of heatmap-based XAI methods, and here we use ODAM [38] as the base XAI method to illustrate and evaluate the effectiveness of our approach. Fig. 1 provides an overview of the framework, detailing how the difference map is constructed and integrated into the visual explanation pipeline.

3.1. Object Discrimination (OD)

OD considers which object was actually detected, identifying the key features that led to detection of the target object, and which were not used by other objects in the scene. While ODAM [38] generates instance-specific heat maps that highlight critical regions influencing each prediction independently, *it is limited in distinguishing their relative*

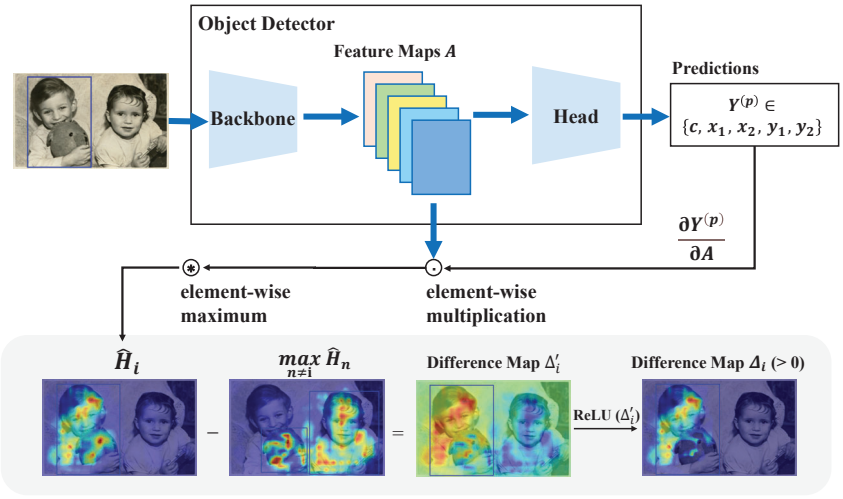


Figure 1. Proposed framework of the difference map for Object Discrimination. The blue box is used only to indicate the target object in the visualization and does not imply that object detections are available at the input stage. The difference map has high values (red) for regions that have larger effect on that object's output compared to other objects, and vice versa for low values (blue). The non-negative entries of the difference map show the features that uniquely led to the detection of that object.

contributions on multiple detected objects. We propose the difference map to address this by differentiating the influence of image regions (pixels) among various objects, providing a more refined explanation for OD.

In ODAM [38], the heat map $H^{(p)}$ for a single detector output $Y^{(p)}$ (e.g., class score, bounding box top-coordinate, etc.) is derived using a pixel-weighted mechanism. Let $A_k \in \mathbb{R}^{H \times W}$ denote the k-th channel of the feature map from a convolutional layer, where H and W are the spatial height and width, and $k \in \{1, \ldots, K\}$ is the channel index. The detector output $Y^{(p)} \in \mathbb{R}$ is a scalar corresponding to the p-th prediction. The importance weight map $w_k^{(p)} \in \mathbb{R}^{H \times W}$ is computed from the gradient map $\frac{\partial Y^{(p)}}{\partial A_k}$ after applying a local smoothing operator $\Phi : \mathbb{R}^{H \times W} \to \mathbb{R}^{H \times W}$:

$$w_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A_k}\right), \tag{1}$$

$$H^{(p)} = \mathrm{ReLU}\left(\sum_{k=1}^{K} w_k^{(p)} \circ A_k\right), \tag{2}$$

where $\circ$ denotes element-wise multiplication. This method highlights regions with strong contributions to $Y^{(p)}$ independently of other outputs.

To provide a holistic explanation of the object prediction, ODAM [38] creates a combined heat map $\hat{H} \in \mathbb{R}^{H \times W}$ from the heat maps for the predicted class and bounding box regression coordinates using an element-wise maximum:

$$\hat{H} = \max\left(H^{(\text{class})}, H^{(x_1)}, H^{(y_1)}, H^{(x_2)}, H^{(y_2)}\right), \tag{3}$$

where $H^{(\text{class})}$ represents the heat map for the object's class prediction, and $H^{(x1)}$, $H^{(y1)}$, $H^{(x2)}$, and $H^{(y2)}$ denote heat maps for its predicted bounding box coordinates.

Min-Max Normalization is then applied to obtain the final heat map:

$$\hat{H} \leftarrow \frac{\hat{H} - \min(\hat{H})}{\max(\hat{H}) - \min(\hat{H})}.$$

In our method, for the ith object instance, we first construct an ODAM combined heat map $\hat{H}_i$, without applying the local smoothing operation Φ. By omitting the smoothing operation, our approach preserves raw gradient information, directly emphasizing pixel contributions. The difference map $\Delta'_i \in \mathbb{R}^{H \times W}$ is then defined to evaluate the feature importance of the ith object instance by comparing its heat map $\hat{H}_i$ with the maximum heat map of other objects $\hat{H}_n$ (for $n \neq i$) in the image:

$$\Delta'_i = \hat{H}_i - \max_{n \neq i} \hat{H}_n. \tag{4}$$

This formulation highlights regions critical to ith object relative to other detected objects, isolating features unique to its detection. To obtain the non-negative difference map that emphasizes the features that are more important to the detected object, we apply the ReLU operation:

$$\Delta_i = \mathrm{ReLU}(\Delta'_i)$$

The non-negative entries of the difference map show the features that uniquely led to the detection of that object.

Min-Max Normalization is then applied to Δ_i or Δ'_i to obtain the final difference maps.

By comparing across multiple objects, our method offers a more granular interpretation of OD, enhancing the visual explanation in complex, crowded scenes. Note that the difference map can be applied to any heatmap-based explanation methods, such as Grad-CAM [27], D-RISE [23], SSGrad-CAM [37] and SSGrad-CAM++ [36]. We provide a comparison of the difference maps generated using different methods in Fig. 4.

4. Experiments

In this section, we conduct experiments on the proposed difference map to: 1) evaluate the explanation method qualitatively and quantitatively; 2) offer a comparison for the results generated with different detectors to provide a better understanding of the characteristics and mechanisms of these detectors; and 3) assess the generality of our approach by comparing difference maps produced using various visual explanation methods.

4.1. Experiment Setup

We implement the methods using a variety of object detection models, including the one-stage detector FCOS [29] and the two-stage detector Faster R-CNN [25] with ResNet-50 [12] as the backbone and FPN [17] as the neck, and the transformer-based detector DETR [2] with ResNet-50 [12] as the backbone, and the fully transformer detector PVT [35] with PVT-Small as the backbone and FPN [17] as the neck. We perform evaluations using the MS-COCO [16] dataset. For qualitative evaluation, we provide examples using the four different detectors, along with a comparison of difference map variants derived from different visual explanation methods [23, 27, 36, 37]. For quantitative evaluation, we compute the Deletion and Insertion metrics [4, 22, 23] on FCOS [29]. All experiments are conducted using PyTorch and an RTX 3090 GPU.

4.2. Qualitative Evaluation

Difference map. The example results of difference maps generated with the four different detectors are presented in Fig. 2. More examples on FCOS [29] are provided in Fig. 3. Compared to ODAM [38], our difference map significantly reduces the highlighting of unrelated or misleading regions, particularly those associated with neighboring objects of the target. For example, in the second row of Fig. 3, the two cows in the image exhibit spatial overlap, and the ODAM heat map shows noticeable energy leakage between them, i.e., regions important to one object are partially attributed to the other. In contrast, our difference map focuses solely on pixels that are specifically important to the target object, with highlighted regions concentrated on the target itself and minimal leakage to adjacent objects.

For the results of object detectors with different architectures in Fig. 2, we have the following observations: 1) The difference maps generated by FCOS [29] and PVT [35] have overall larger areas of activation for each object instance, with a deeper coloration implying higher activation values; 2) The highlighted regions in the difference maps generated by Faster R-CNN [25] are confined within the detection boxes and exhibit a rectangular shape. This feature may be attributed to the RPN stage of Faster R-CNN [25], which produces region proposals that are subsequently passed to the second stage for classification; 3) The regions highlighted in the difference maps generated by DETR [2] are sparser compared to the other two detectors. Additionally, these regions tend to concentrate around the object's edges, which is likely due to the nature of the DETR [2] decoder, which aims to recover the bounding position from the decoder's input queries.

In the difference maps of the second example (bottom row containing a person, horse, and dog), it is interesting to note that both FCOS [29] and DETR [2] attribute the legs of the person to the person detection (legs are highlighted red for the person bounding box). On the other hand, for Faster R-CNN [25], the legs are attributed equally to the horse and the person (the legs are neutral green color). Thus, we may infer that for Faster R-CNN, the legs at the top of the horse are equally important for both the horse and person.

Difference map generated using other methods. Our proposed difference map is generalizable as it can be integrated with any heatmap-based visual explanation method. In Fig. 4, we present a comparative analysis of difference maps generated by our method, which builds upon ODAM [38], alongside difference maps generated using several established heatmap-based methods, including Grad-CAM [27], D-RISE [23], SSGrad-CAM [37] and SSGrad-CAM++ [36] on FCOS [29]. Our qualitative results demonstrate that augmenting these methods with the difference map refines the original heat maps by effectively suppressing the highlighting of irrelevant or misleading regions associated with detected objects. Notably, the difference map variant of D-RISE [23] significantly reduces noisy areas, leading to clearer and more focused explanations. These results suggest that our method can enhance existing visual explanation methods by improving both interpretability and user-friendliness.

4.3. Quantitative Evaluation

Deletion and Insertion. Deletion and Insertion metrics [4, 22, 23] evaluate changes in a detector's outputs by selectively deleting or inserting pixels from the original image based on their importance. In the Deletion process, pixels are replaced with random values in descending order of importance—starting from the most important ones in the heat map (typically shown in red). The corresponding de-

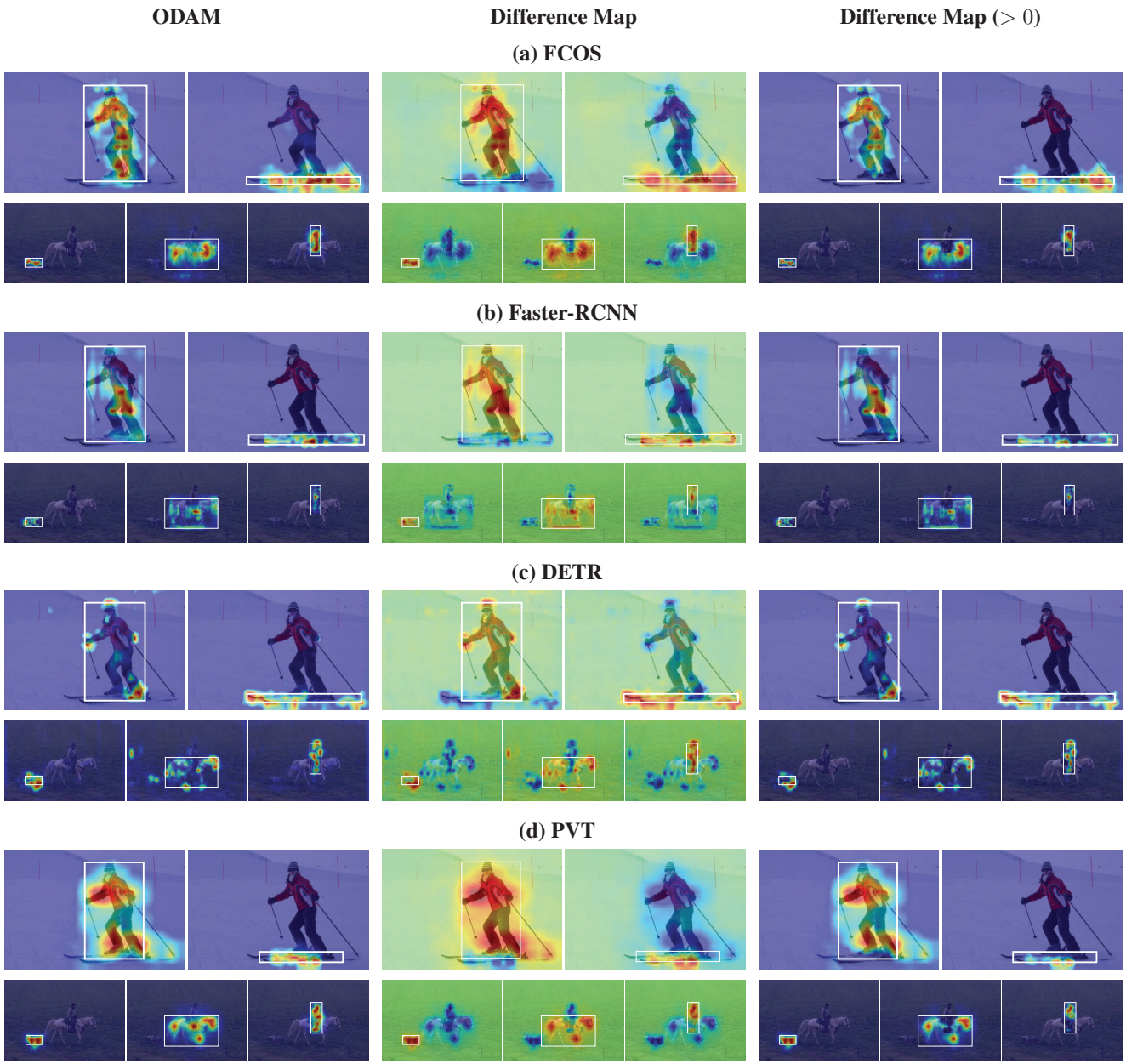


Figure 2. Comparison of ODAM saliency maps and our difference maps across multiple detectors: (a) FCOS, (b) Faster-RCNN, (c) DETR, (d) PVT. We also visualize the positive part of the difference map (denoted as "> 0") to intuitively visualize the new information it uncovers. The bounding box indicates the target object. Red highlights in the difference map (middle) indicate influence only to the target object, while blue highlights indicate influence to only other objects.

crease in the detector's confidence score is then measured as a percentage. Conversely, the Insertion process incrementally adds pixels back to a blank image, beginning with the most important ones, and tracks the increase in confidence scores to assess the contribution of each region. The underlying rationale is that important pixels should have a greater influence on the target prediction and thus lead to more significant changes in the model's output when altered. In the experiments, we divide the total area of the ground truth bounding box of the target object into 10 equal steps, each representing 10% of the total area, and record the results for 10 steps of Deletion and Insertion. Note that pixels can

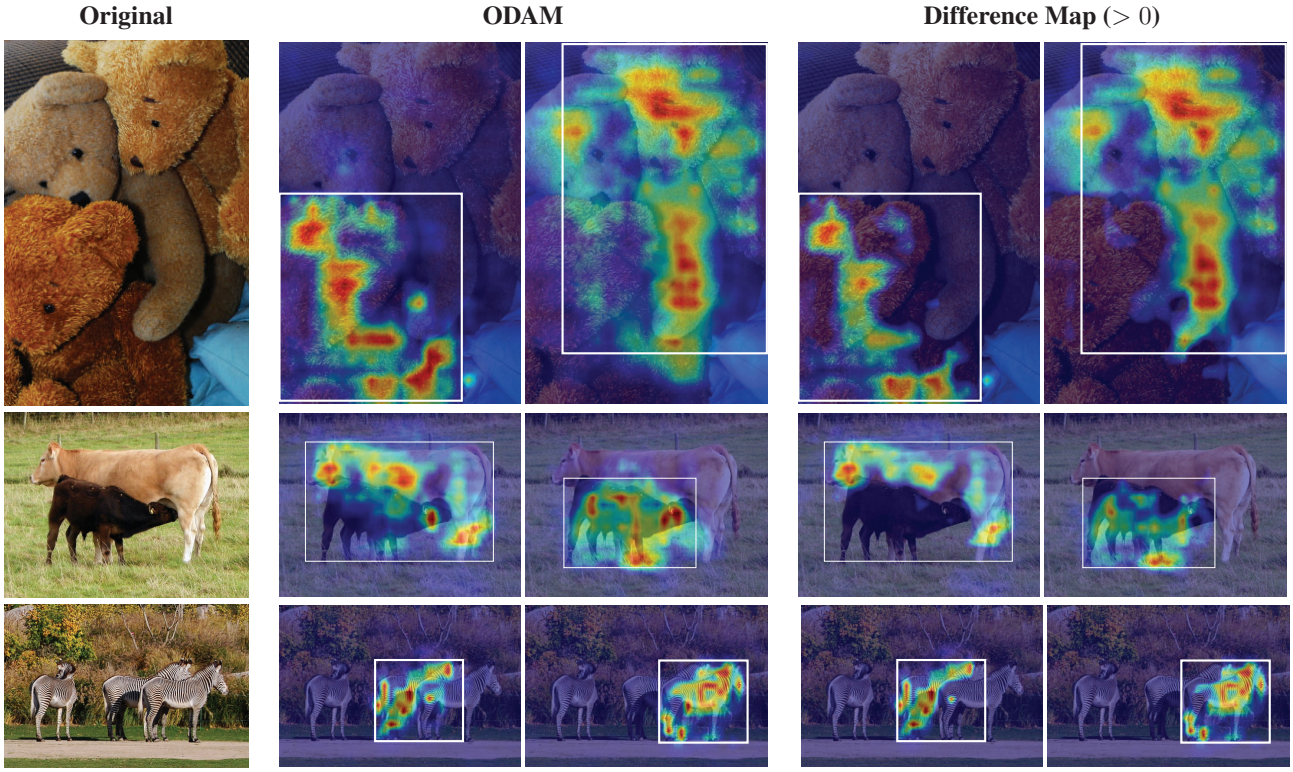


Figure 3. Comparison of original images, ODAM, and our difference maps for three different examples using FCOS. We vsualize the positive part of the difference map (denoted as "> 0") to intuitively show the information it uncovers compared to ODAM.

Method	Target Object		Other Objects	
	Del (↓)	Ins (↑)	Del (↑)	Ins (↓)
Grad-CAM	94.88	9.40	89.54	7.23
Grad-CAM (Diff.)	93.76	7.84	91.01	5.99
SS-GradCAM	64.91	25.67	94.99	2.66
SS-GradCAM (Diff.)	65.24	26.33	95.70	3.10
SS-GradCAM++	<u>60.84</u>	26.86	96.27	2.68
SS-GradCAM++ (Diff.)	61.40	37.16	<u>98.44</u>	<u>1.94</u>
ODAM	61.27	**46.06**	96.12	3.14
Difference Map (ours)	**60.74**	<u>44.25</u>	**99.01**	**1.58**

Table 1. AUC for the Deletion and Insertion curves in Figure 5. "Diff" indicates the Diference map calculated using the corresponding visual explanation method. In the Deletion process, a lower AUC for the *target object* indicates that removing important pixels significantly reduces its confidence score, while a higher AUC for *other objects* indicates that their predictions remain largely unaffected. Conversely, in the Insertion process, a higher AUC for the *target* and lower AUC for *others* reflects effective localization of discriminative regions. The best results are shown in bold, and the second-best are underlined.

be deleted/inserted anywhere in the image according to the heat map, and is not restricted to just the bounding box.

Experiment Setup. For the difference map, we define each detected object in the image as the target object, with the remaining objects in the same image as other objects. We conduct the Deletion and Insertion experiment for the difference map using FCOS [29] with the ResNet-50 backbone [12]. We calculate the Deletion and Insertion metrics for predictions with the class score > 0.4 and IoU > 0.5 for images in the COCO validation set that contain multiple such predictions, as the difference map is designed to distinguish one object from multiple objects. For the target object, we record its own score change, while for other objects, we record the change in their average score. Ideally, the confidence score of the target object should exhibit a significant change during Deletion or Insertion, indicating that the identified regions are indeed critical for its detection. In contrast, the scores for other objects should remain relatively stable, as the difference map is specifically designed to highlight features that are uniquely important to the target object, rather than shared or overlapping features. This behavior reflects strong object discrimination and demonstrates the ability of our method to isolate object-specific information in complex scenes.

Results. Fig. 5 compares the score changes at each step of the Deletion and Insertion experiments across various visual explanation methods and our proposed difference map. The evaluated methods include the original ODAM [38],

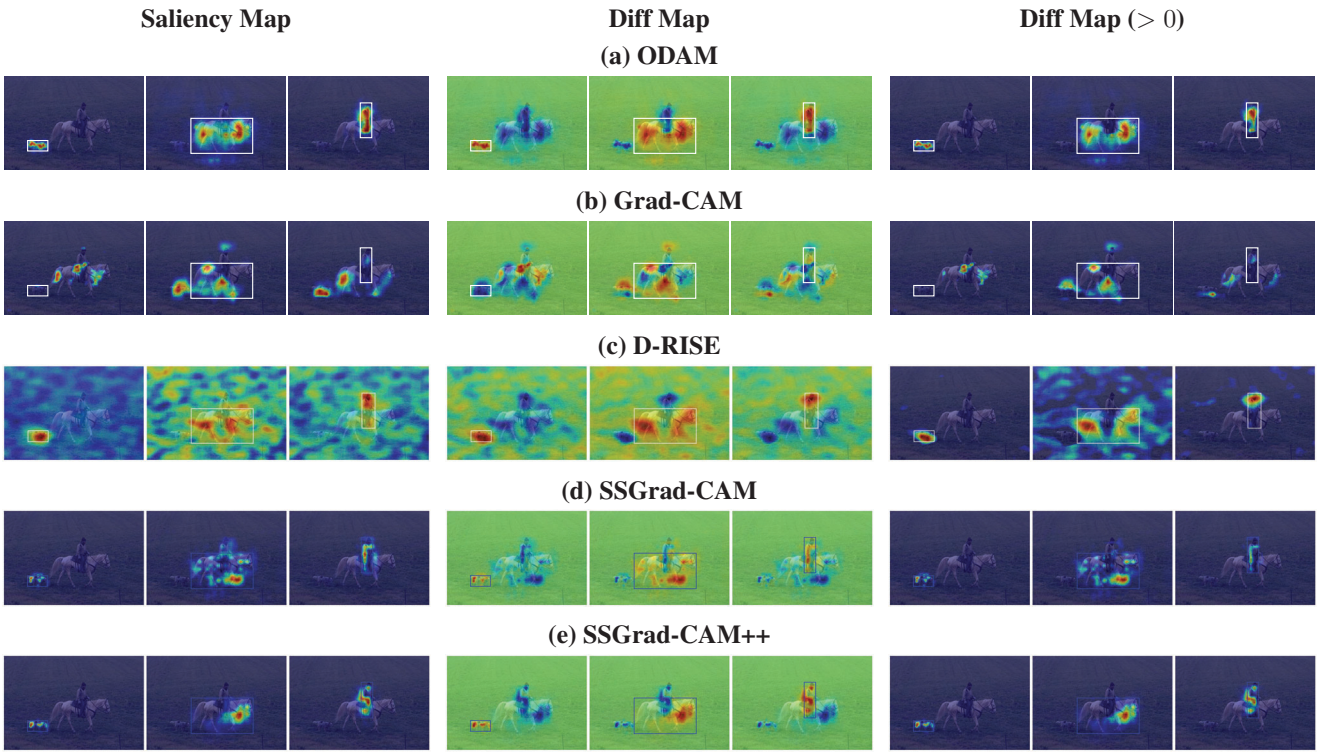
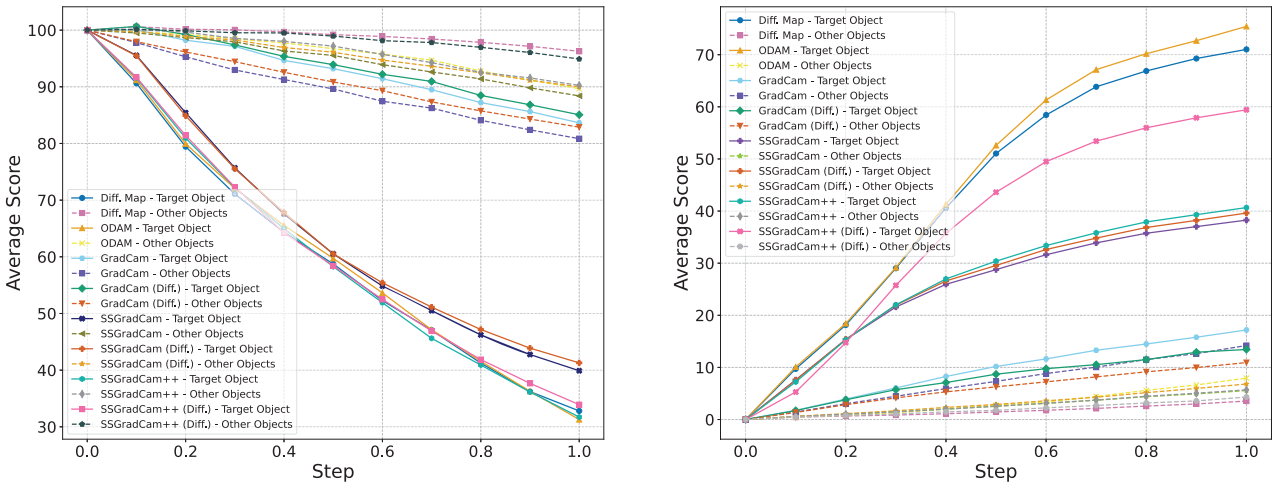


Figure 4. Comparison of saliency maps and difference maps generated using our method based on: (a) ODAM, (b) Grad-CAM, (c) D-RISE, (d) SSGrad-CAM, and (e) SSGrad-CAM++ on FCOS. Left: The standard saliency map. Middle: Difference map including negative values. Right: Difference map with only positive values. The difference map variants improves the original heat maps by effectively reducing the highlighting of irrelevant regions of detected objects, leading to clearer and more focused explanations.

as well as the original versions and corresponding difference map variants (denoted as "Diff.") of Grad-CAM [27], SSGrad-CAM [37] and SSGrad-CAM++ [36]. Our method demonstrates substantial score changes for the "target object", while exhibiting relatively minor score changes for "other objects" at each step compared to other methods. This indicates the effectiveness of the difference map, verifying that regions highlighted in the difference map have a more significant impact on the target object compared to other objects.

Tab. 1 shows the comparison of the corresponding area under the curve (AUC) for the Deletion and Insertion curves. In the Deletion process, a lower AUC for the *target object* is desirable, indicating that the removal of important pixels results in substantial drops in confidence scores. In contrast, a higher AUC for the *other objects* is preferable, suggesting that their scores remain largely unaffected and the highlighted regions are indeed specific to the target object. Conversely, in the Insertion process, the interpretation is reversed: a higher AUC for the *target object* and a lower AUC for *other objects* are indicative of successful localization of discriminative regions. The AUC values for other objects obtained using our difference map are the best among all methods in both the Deletion and Insertion processes, while the AUC value for the target object is also the best in the Deletion process. This indicates that the highlighted regions produced by the difference map effectively discriminate key features specific to the target object, capturing the relative importance of each pixel across different predictions within the same scene.

The difference map variants of the original heatmap-based explanation methods demonstrate slightly superior AUC results. Compared to the original ODAM, the ODAM difference map (ours) achieves better AUCs in the Deletion and Insertion processes for other objects, as well as in the Deletion process for the target object. While its AUC for the target object in the Insertion process (44.25%) is slightly lower than that of ODAM (46.06%), it still surpasses all other methods. This slight drop is attributed to the more focused highlighting of the target object in the difference map, which results in a smaller highlighted area. Regarding other heat-map methods, the difference map version of SSGrad-CAM++ achieves improved AUCs for other objects in both Deletion and Insertion processes, along with



(a) (b)

Figure 5. Comparison of average class prediction score vs. (a) Deletion steps and (b) Insertion steps for ODAM and the difference map on FCOS. For the "target object", lower scores are better for deletion, while higher scores are better for insertion. For "other objects", higher scores are better for deletion, and lower scores are better for insertion. In both ODAM and the difference map, the *target object*'s score exhibits substantial changes. However, for *other objects*, the difference map leads to smaller score change compared to ODAM, demonstrating better ability to find features uniquely important for the target object.

a notable 10.3% increase in the Insertion process for the target object. Similarly, the SSGrad-CAM variant shows modest gains in AUC for the target object in the Insertion process and for other objects in the Deletion process. The class-specific method Grad-CAM performs poorly in isolating instance-level objects, resulting in the worst AUCs. However, its difference map variant improves AUCs in both Deletion and Insertion processes for other objects, as well as in the Deletion process for the target object. These results highlight the effectiveness of the proposed difference map in improving existing visual explanation methods.

5. Discussions and Conclusion

This paper presents the difference map, an effective visual explanation method designed to enhance the interpretability of object detection models by focusing on Object Discrimination (OD). Object Discrimination, as measured by our difference map approach, identifies regions (key features) that are more critical for the target object compared to other objects in the same scene. This is achieved by refining the instance-specific heatmaps produced by ODAM, thereby offering a more detailed and object-specific understanding of feature importance. To validate the effectiveness and generality of our method, we conduct extensive qualitative and quantitative evaluations across various types of object detectors, including one-stage, two-stage, and transformer-based models. Furthermore, we demonstrate that our difference map can be integrated into a range of existing heatmap-based visual explanation methods, improving their ability to generate focused and interpretable visualizations. These results collectively highlight the versatility and utility of the proposed method in advancing the transparency of object detection.

Limitations and future work. One aspect that the current explanation framework does not consider is Object-Object Discrimination, i.e., why 2 neighboring objects are detected as separate objects and not as one. We will consider this in future work, focusing on the bounding box information to look into the boundary between neighboring objects and explain through the boundary's influence. Another limitation is that the proposed methods do not enhance the model's performance, although that is not the focus of this paper. For difference maps, we will investigate how to utilize it in knowledge distillation to improve model's accuracy, e.g., borrowing ideas from [39].

6. Acknowledgement

This work was supported by a Strategic Research Grant from City University of Hong Kong (Project No. 7005995).

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 9525–9536, 2018. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas

Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, 2020. 2, 4

[3] Chun-Hao Kingsley Chang, Elliot Creager, Anna Goldenberg, and David Kristjanson Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2018. 1

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018. 1, 2, 4

[5] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. page 6970–6979, 2017. 1

[6] Saurabh Desai and Harish G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 972–980, 2020. 1

[7] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, page 590–601, 2018. 2

[8] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[9] Ross Girshick. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 2

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation . In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 1, 2

[11] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2376–2384, 2019. 2

[12] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 4, 6

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation Networks for Object Detection . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3588–3597, 2018. 1

[15] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. BBAM: Bounding Box Attribution Map for Weakly Supervised Semantic and Instance Segmentation . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2643–2651, 2021. 1

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, 2014. 4

[17] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2017. 4

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection . In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017. 2

[19] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, 2019. 2

[20] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, 2017. 1

[21] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Karthikeyan Shanmugam, and Chun-Chen Tu. Generating contrastive explanations with monotonic attribute functions. *ArXiv*, abs/1905.12698, 2019. 2

[22] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151, 2018. 4

[23] Vitali Petsiuk, R. Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11438–11447, 2020. 1, 2, 4

[24] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2015. 2

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06):1137–1149, 2017. 2, 4

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, 2016. 1

[27] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. 1, 2, 4, 7

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 1

[29] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9626–9635, 2019. 2, 4, 6

[30] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*, page 650–665, 2021. 2

[31] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017. 2

[32] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and Fine-Grained Visual Explanations for Convolutional Neural Networks . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9089–9099, 2019. 1

[33] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2020. 1

[34] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8978–8987, 2020. 2

[35] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021. 2, 4

[36] Toshinori Yamauchi. Spatial sensitive grad-cam++: Improved visual explanation for object detectors via weighted combination of gradient map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8164–8168, 2024. 1, 2, 4, 7

[37] Toshinori Yamauchi and Masayoshi Ishikawa. Spatial sensitive grad-cam: Visual explanations for object detection by incorporating spatial sensitivity. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 256–260, 2022. 1, 2, 4, 7

[38] Chenyang Zhao and Antoni B. Chan. Odam: Gradient-based instance-specific visual explanations for object detection. In *ICLR*, 2023. 1, 2, 3, 4, 6

[39] Chenyang Zhao, Janet H. Hsiao, and Antoni B. Chan. Gradient-based instance-specific visual explanations for object specification and object discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9): 5967–5985, 2024. 8

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization . In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1