Gradient-based Instance-Specific Visual Explanations for Object Specification and Object Discrimination

Chenyang Zhao, Janet H. Hsiao, and Antoni B. Chan

Abstract—We propose the gradient-weighted Object Detector Activation Maps (ODAM), a visual explanation technique for interpreting the predictions of object detectors. Utilizing the gradients of detector targets flowing into the intermediate feature maps, ODAM produces heat maps that show the influence of regions on the detector's decision for each predicted attribute. Compared to previous works on classification activation maps (CAM), ODAM generates instance-specific explanations rather than class-specific ones. We show that ODAM is applicable to one-stage, two-stage, and transformer-based detectors with different types of detector backbones and heads, and produces higher-quality visual explanations than the state-of-the-art in terms of both effectiveness and efficiency. We discuss two explanation tasks for object detection: 1) object specification: what is the important region for the prediction? 2) object discrimination: which object is detected? Aiming at these two aspects, we present a detailed analysis of the visual explanations of detectors and carry out extensive experiments to validate the effectiveness of the proposed ODAM. Furthermore, we investigate user trust on the explanation maps, how well the visual explanations of object detectors agrees with human explanations, as measured through human eye gaze, and whether this agreement is related with user trust. Finally, we also propose two applications, ODAM-KD and ODAM-NMS, based on these two abilities of ODAM. ODAM-KD utilizes the object specification of ODAM to generate top-down attention for key predictions and instruct the knowledge distillation of object detection. ODAM-NMS considers the location of the model's explanation for each prediction to distinguish the duplicate detected objects. A training scheme, ODAM-Train, is proposed to improve the quality on object discrimination, and help with ODAM-NMS. The code of ODAM is available: https://github.com/Cyang-Zhao/ODAM.

Index Terms—gradient-based explanation, instance-level explanation, object specification, object discrimination, explaining object detection, human eye gaze, explainable AI, non-maximum suppression, knowledge distillation, deep learning

1 Introduction

Significant breakthroughs have been made in object detection and other computer vision tasks due to the development of deep neural networks (DNN) [1]. However, the unintuitive and opaque process of DNNs makes them hard to interpret. As spatial convolution is a frequent component of state-of-the-art models for vision tasks, class-specific attention has emerged to interpret CNNs, which has been used to identify failure modes [2, 3], debug models [4] and establish appropriate users' confidence about models [5]. These explanation approaches produce heat maps locating the regions in the input images that the model looked at, representing the influence of different pixels on the model's decision. Gradient visualization [6], Perturbation [7], and Class Activation Map (CAM) [8] are three widely adopted methods to generate the visual explanation map. However, these methods have primarily focused on image classification [9, 10, 5, 11, 12, 13], or its variants, e.g., visual question answering [14], video captioning [15, 16], and video activity recognition [16].

Generating explanation heat maps for object detectors is an under-explored area. The first work in this area is D-RISE [17], which extends RISE [9] for explaining image classifiers to object detectors. As a perturbation-based approach, D-RISE first randomly generates a large number of binary masks, resizes them to the image size, and then perturbs the original input to observe the change in the model's prediction. However, the large number of inference calculations makes the D-RISE computationally intensive, and the quality of the heat maps is influenced by the dataset type and mask resolution (e.g., see the 4th row of Fig. 1). Furthermore, D-RISE only generates an overall heat map for the predicted object, which is unable to show the influence of regions on the specific attributes of a prediction, e.g., class probability and regressed bounding box corner coordinates.

The popular CAM-based methods for image classification are not directly applicable to object detectors. CAM methods generate heat maps for classification via a linear combination of the weights and the activation maps, such as the popular Grad-CAM [5] and its variants. However, for object detection task, Grad-CAM and Grad-CAM++ generate explanation maps with poor localization ability and tend to provide class-specific explanations. As a result, Grad-CAM produces heat maps that highlight multiple objects of a category instead of explaining a single detection (e.g., see the 2nd and 3rd rows of Fig. 1). For object detection, the explanations should be *instance-specific* rather than class-specific, so as to discriminate each individual object. Exploring the spatial importance of different objects can help interpret the models' decision and show the important area in the feature maps for each prediction.

Considering that direct application of existing CAM methods to object detectors is infeasible and the drawbacks of the current

• *Chenyang Zhao, and Antoni B. Chan (corresponding author) are with the Department of Computer Science, City University of Hong Kong. Janet H. Hsiao is with the Department of Psychology, University of Hong Kong. E-mail: zhaocy2333@gmail.com, abchan@cityu.edu.hk.*

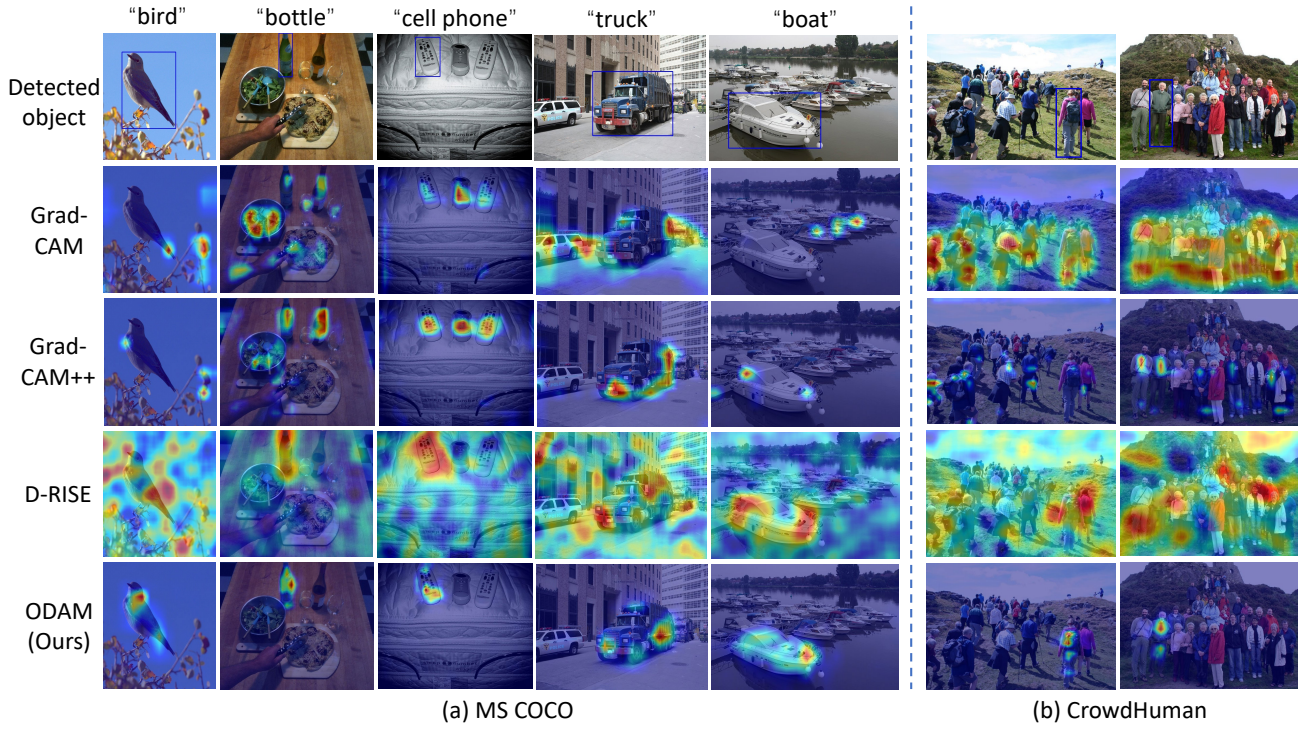


Fig. 1: Comparison of heat maps from Grad-CAM [5], Grad-CAM++ [11] D-RISE [17] and our ODAM for interpreting predictions from the FCOS [18] detector. Visual explanations are provided for predicted instances of different classes on *MS COCO val* set and single person class on *CrowdHuman val* set, and the blue boxes show the corresponding detected objects. Grad-CAM and Grad-CAM++ generate low-quality explanations for object detection and tend to highlight multiple objects of the same category instead of the specific object instance. D-RISE maps have noisy backgrounds and its effectiveness depends on the mask size; In the examples, with the 16x16 mask it performs better for smaller objects (such as "bottle" and "cell phone") than larger objects (such as "bird" and "truck"). Moreover, D-RISE has poor performance on the CrowdHuman dataset with one-class crowded scenes. In the last row, our ODAM generalizes well on different scenes, and generates instance-specific heat maps with less noise and is robust to object size.

state-of-the-art D-RISE, we propose gradient-weighted *Object Detector Activation Maps* (ODAM). ODAM adopts a similar assumption as Grad-CAM that feature maps correlate with some concept for making the final outputs. Thus ODAM uses the gradients w.r.t. each pixel in the feature map to obtain the explanation heat map for each attribute of the object prediction. Compared with the perturbation-based D-RISE, ODAM is more efficient and generates less noisy heat maps (see the 5^{th} row of Fig. 1), while also explaining each prediction separately.

Object detectors are different from image classifiers in that detectors perform classification and localization of multiple object instances within the image, whereas image classifiers only predict the class of the main (usually single) object. Thus for object detectors, we consider two types of explanation tasks: 1) "object specification"; 2) "object discrimination". *Object specification* is the traditional explanation task, which aims to answer "what context/features are important for the prediction?", via a heat map that highlights the important regions for the final prediction. *Object discrimination* is an explanation task to answer "which object was actually detected?", which is a unique task for object detection where there are multiple classified objects. For the discrimination task, the explanation map is expected to show which instance was considered when the model made the prediction.

In the experiments, we consider both qualitative evaluation by visualization of explanation maps and quantitative evaluation of ODAM on both object specification and object discrimination, and demonstrate that ODAM outperforms current methods. Moreover, we carry out user trust studies with visual explanations on both object specification and discrimination. In particular, we collect human eye gaze data (i.e., human attention data) during an object explanation task, and investigate how well the visual explanations of object detectors agrees with human explanations, and whether this agreement is related with user trust of the detector. Our experiments show that ODAM generates heat maps that are more similar to human attention, while also being more trustworthy.

Finally, in this paper, we also provide insight into how to use the visual explanation of object specification and object discrimination to boost detection performance in particular regimes. For object specification, we propose ODAM-based knowledge distillation (ODAM-KD) for object detection, which distills on the crucial regions emphasized by the ODAM maps of the key predictions, and a key prediction selection module is designed for choosing the important instances. For object discrimination, we propose ODAM-NMS, which uses the instance-level heat maps from ODAM to aid the non-maximum suppression (NMS) process of removing duplicate and preserving overlapped predictions in the crowded scenarios. To guide the explanation map more towards object discrimination, we propose a training scheme, ODAM-Train, which introduces consistency and separation losses to encourage the model to produce consistent heat maps for the same object, and distinctive heat maps for different objects, thus making the heat map more discriminative. Experiments show that both ODAM-KD and ODAM-NMS are effective.

In summary, the contributions of this paper are:

1) We investigate ODAM, a gradient-based visual explanation approach to produce instance-specific heat maps for explain-

ing prediction attributes of object detectors, which is more efficient and robust than the current state-of-the-art.
2) We demonstrate the generalizability of ODAM by exhibiting explanations on one-stage, two-stage, and transformer detectors with different types of backbones and detector heads.
3) Besides the traditional explanation task, object specification, we explore a unique explanation task for detector, object discrimination, for explaining which object was detected. Comprehensive experiments are conducted to verify the ability of ODAM on both the two tasks.
4) We conduct user trust studies on visual explanations, and investigate how well the visual explanations of object detectors agrees with human explanations, via human eye gaze.
5) Based on the object specification ability of ODAM, We propose ODAM-KD, where the student detector learns features from the teacher detector based on the attention map generated by ODAM with the selected key instances.
6) Based on the object discrimination ability of ODAM, we propose ODAM-NMS, which uses the instance-level heat maps generated by ODAM with ODAM-Train to remove duplicate and preserve overlapped predictions during NMS.

A preliminary version of ODAM appears in [19]. The extensions over the conference version are as the follows. First, by conducting user trust studies and investigating the similarity between human attention and XAI heat maps, we discover that XAI heat maps that are more trustworthy also have higher similarity to human attention. Second, we enrich and complete the qualitative evaluation of ODAM by adding: (1) the visualization of ODAM with different feature map layers; (2) the analysis of error modes using ODAM; (3) a comparison with the attention maps of DETR; (4) application on the multi-modal visual grounding model GLIP. Finally, we propose a new application of ODAM, ODAM-KD, which applies the object specification ability of ODAM to the knowledge distillation task on object detection.

The remainder of this paper is organized as follows. The related works of object detection, explanation by visualization, knowledge distillation, and advanced NMS are briefly reviewed in §2. ODAM is introduced in §3, and the experiment results and analysis of ODAM are presented in §4. We then introduce two applications: ODAM-KD with object specification in §5; ODAM-NMS and ODAM-Train with object discrimination in §6.

2 Related Works

We first briefly review main stream object detectors, and then discuss related visual explanation algorithms. Finally, the related methods for knowledge distillation and NMS are introduced.

2.1 Object detection

Object detectors are generally composed of a backbone, neck and head. Based on the type of head, detectors can be mainly divided into one-stage and two-stage methods. Two-stage approaches perform two steps: generating region candidates (proposals) and then using RoI (Region of Interest) features for the subsequent object classification and location regression. The representative two-stage works are the R-CNN family, including R-CNN [20], Fast R-CNN [21], Faster-RCNN [22], and Mask R-CNN [23]. One-stage methods remove the RoI feature extraction and directly perform classification and regression on the entire feature map, and typical methods are YOLO [24], RetinaNet [25], and FCOS [18]. Recently transformers are successfully applied to the detector structure, being utilized as a backbone to extract features from image (e.g. pyramid vision transformer (PVT) [26]) or as the detector head (e.g. DETR [27]). Our ODAM can generate heat maps for both one- and two-stage detectors with no limitation of the types of backbones and heads, as well as transformer-based detectors. We mainly adopt Faster R-CNN and FCOS in our experiments.

2.2 Explanation by visualization

Since visualizing the importance of input features is a straightforward approach to interpret a model, many works visualize the internal representations of image classifier CNNs with heat maps. Gradient visualization methods [6] backpropagate the gradient of a target class score to the input image to highlight the "important" pixels, and other works [28, 29, 30, 31, 32] manipulate this gradient or use a set of purposely designed propagation rules [33, 34] to improve the results qualitatively. These visualizations are fine-grained but not class-specific. [35] propose a new back-propagation scheme for generating task-specific attention maps. [36] proposed a novel way to compute relevancy for Transformer networks specifically.

Perturbation-based methods [9, 7, 37, 38, 39, 10, 40, 41] perturb the original input and observe the changes in output scores to determine the importance of regions. Most black-box methods are intuitive and highly generalizable, but computationally intensive. Furthermore, the type or resolution of the perturbation greatly influences the quality of visualization results.

CAM-based explanations, e.g. CAM [8], Grad-CAM [5], Grad-CAM++ [11], and related works [42, 43, 44] produce a heat map from a selected intermediate layer by linearly combining its feature activation maps with weights that indicate each feature's importance. For example, Grad-CAM defines the weights as the global average pooling of the corresponding gradient map, computed using back-propagation. Some gradient-free CAMs [45, 12, 13] adopt the perturbation to generate weights from class score changes.

Although Grad-CAM has been adopted to study adversarial context patches in single-shot object detectors [46], the explanations are still category-specific. [17] describes the reasons that make direct application of existing classifier explanation methods infeasible for object detector, and then proposes D-RISE [9], a black-box perturbation-based method. Hence, D-RISE inherits the pros and cons of RISE: high-generalizability due to the black-box nature, but time-consuming and noisy due to the inference procedure. To explore white-box explanations of detectors, we propose ODAM, which uses gradient information to generate importance heat map for instance-specific detector explanations.

2.3 Knowledge Distillation

Knowledge distillation is first proposed by Hinton et al. [47] for the classification task. As an effective method for model compression and accuracy boosting, KD has also been applied to image classification [48, 49], as well as semantic segmentation [50], face recognition [51], pretrained language model [52], etc.

KD for object detection has also be considered recently. Chen et al. [53] first propose distillation on all components of detector (the neck feature, classification and regression heads), but treating all features and proposals equally introduces much noise and leads to a suboptimal result. Thus previous methods aim to select and weigh the important regions on a feature map,

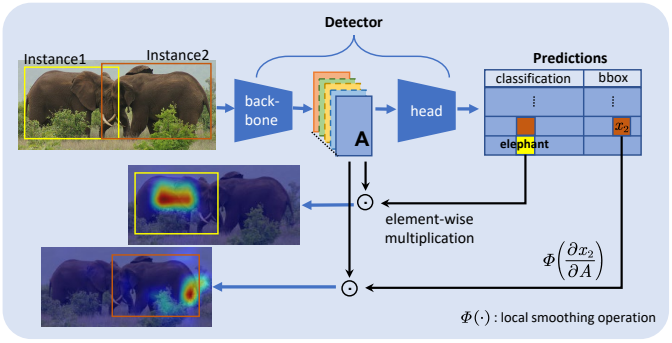


Fig. 2: ODAM framework. ODAM generates instance-specific heat maps for explaining attributes in predictions of an object detector.

which may highly influence the distillation effect. Li et al. [54] choose the features sampled from RPN, and GID [55] distills the areas where generate distinctive predictions. Wang et al. [56] and Sun et al. [57] both utilize designed region masks generated from ground-truth bounding boxes for distillation. Guo et al. [58] decouple the foreground and background and distill them with different attentions. FKD [59] and FGD [60] utilize the bottom-up attentions as mask to guide the distillation, and add Non-local Module [61] and GcBlock [62] to distill the relation of pixels.

Compared with the attention mask designs in previous works, we propose ODAM-KD, which utilizes ODAM to generate *top-down* feature attention based on the student and teacher predictions during training. The key predictions that cause the performance gap between student and teacher are selected, and their ODAM explanation maps are generated as the distillation attention mask.

2.4 Advanced NMS

Classic NMS assumes that multiple instances rarely overlap, and thus high IoU (intersection over union) of two bounding boxes indicates duplicate detections. However, high IoU will occur between objects in crowded scenes, resulting in erroneous removal of instances. Thus, more advanced NMS methods are proposed to mitigate the over-dependence on IoU. SoftNMS [63] decreases the scores of duplicates according to the IoU instead of removing overlapping predictions. AdaptiveNMS [64] adjusts the NMS threshold based on the predicted local object density for each prediction. Relation Network [65] adds a relation module to the detection network and uses it to learn NMS inside the network. These methods either use extra predicted cues, but still assume that high IoU correspond to duplicate detections, or modify the detectors to be more complex. Relying on IoU is not enough in crowded scenes where objects partially occlude each other, and thus their IoUs are naturally large. In these cases, internal information about the predictions is required. FeatureNMS [66] encodes features for predictions, and trains their distances between the same object to be smaller than those of different objects. In contrast, we propose ODAM-NMS, which uses the correlations between instance-level explanation maps and the their box IoUs to remove duplicate proposals of the same object. Compared with [66], our ODAM-NMS is more stable and can also be interpreted to explain which objects were detected (*i.e.*, object discrimination).

3 ODAM: Object Detection Activation Map

Given an image $\mathcal{I}$, the detector model outputs multiple predictions, with each prediction p consisting of the class score $s_c^{(p)}$ and bounding box $B^{(p)} = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$. Our goal is to generate heat maps to indicate the important regions that have a positive influence on the output of each prediction.

In Grad-CAM [5] and its generalization Grad-CAM++ [11], the final score for a particular class Y_c is predicted from the whole image and the algorithm ignores distinguishing object instances within. Their explanation starts from the assumption that the score to be interpreted can be written as a linear combination of its global pooled last convolutional layer feature maps $\{A_k\}_k$, $Y_c = \sum_k w_k^c \sum_{ij} A_{ijk} = \sum_{ij} \sum_k w_k^c A_{ijk}$, where A_{ijk} indexes location (i, j) of A_k. Thus, the class-specific heat map $H_{ij}^c = \sum_k w_k^c A_{ijk}$ summarizes the feature maps with w_k^c capturing the importance of the k-th feature. To obtain the importance of each feature channel, Grad-CAM estimates w_k^c by global average pooling the gradient map $\partial Y_c / \partial A_k$, while Grad-CAM++ uses a weighted global pooling. However, both methods are limited to class-specific explanations, due to computing a channel-wise importance, which ignores the spatial information that is essential for interpreting different object instances.

Based on the above analysis, we assume that any predicted object attribute scalar $Y^{(p)}$ of a particular instance p can be written as a linear element-wise weighted combination of the feature map, and then the instance-specified by summarizing the feature maps with weight $w_{ijk}^{(p)}$ that captures the importance of *each pixel and each channel*,

$$Y^{(p)} = \sum\nolimits_k \sum\nolimits_{ij} w_{ijk}^{(p)} A_{ijk}, \quad H_{ij}^{(p)} = \sum\nolimits_k w_{ijk}^{(p)} A_{ijk}. \quad (1)$$

Here $Y^{(p)}$ could be the classification score or a predicted bounding box coordinate, depending on the desired explanation.

Previous gradient-based works [6, 28, 5, 11] have shown that the partial derivative w.r.t. A_{ijk} can reflect the influence of the k-th feature at (i, j) on the final output. Thus, we set the importance weight map $W_k^{(p)} = [W_{ijk}^{(p)}]_{ij}$ according to the gradient map $\partial Y^{(p)} / \partial A_k$ after a local smoothing operation Φ, and the corresponding heat map for scalar output $Y^{(p)}$ is obtained through a pixel-weighted mechanism:

$$W_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A_k}\right), \quad H^{(p)} = \mathrm{ReLU}\left(\sum\nolimits_k W_k^{(p)} \circ A_k\right), \quad (2)$$

where $\circ$ is element-wise multiplication. In the qualitative experiments to produce better visualizations, we adopt convolution with a Gaussian kernel as the smoothing operation Φ, and adaptively decide the size of the kernel based on the size of the predicted object in the feature map. In the quantitative experiments, the local smoothing is not used, i.e., Φ is the identity function, in order to maintain the best spatial accuracy. Finally, we note that Grad-CAM is a special case of ODAM when the smoothing function Φ is set to the global average pooling operation. Fig. 2 shows our ODAM framework.

When the scalar output $Y^{(p)}$ is a class score, ODAM highlights the important feature regions used by the detector to classify the instance p. Note that $Y^{(p)}$ could be any differentiable attribute of the predictions. For example, in our experiments we also examine the heat maps related to the predicted coordinates of the regressed bounding box (see Fig. 3).

Our work is the first analysis and successful attempt to use a white-box method to generate instance-level explanations for

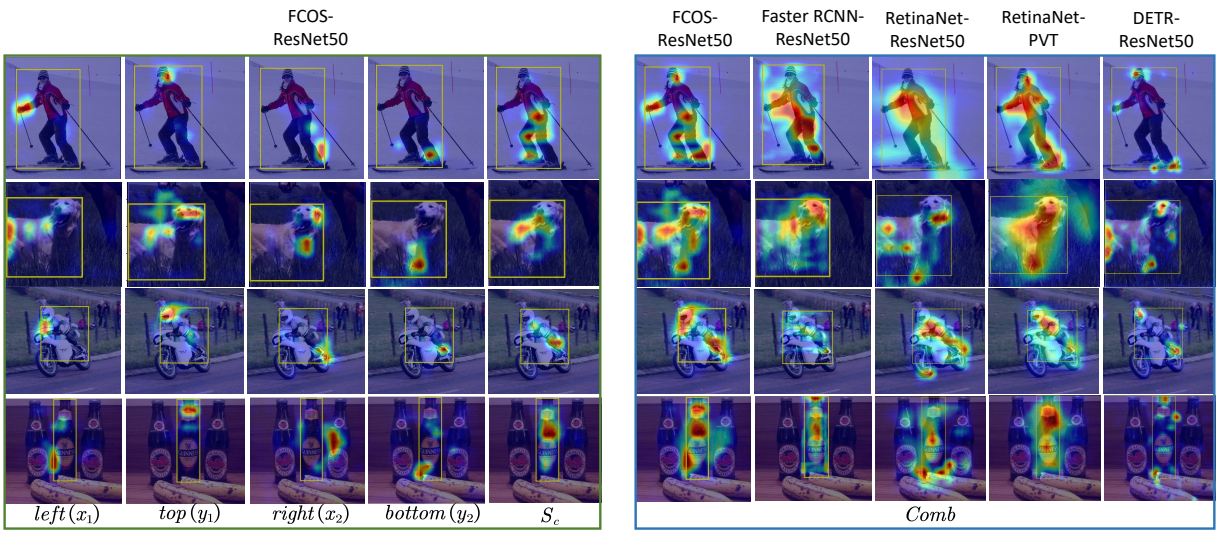


Fig. 3: Heat map explanations computed from different detectors with ResNet50 and pyramid vision transformer (PVT) [26] as backbones. (left) The heat maps explain important regions for each predicted attribute (class score s_c and bbox coordinates x_1, y_1, x_2, y_2) from FCOS. In FCOS, the bbox coordinates are relative to the anchor center, with positive values indicating a larger box, and thus the highlighted regions for the bbox coordinates are important for expanding the bbox. (right) The combined heat maps for the entire predictions for one-stage RetinaNet, FCOS, the two-stage Faster R-CNN, and transformer-based DETR [27]. Features from the last stage of ResNet50 are used to explain DETR because there is no detector neck, while features from Feature Pyramid Network (FPN) [69], the detector neck, are used for other methods.

object detector predictions, rather than only class-level explanations. Using PyTorch, gradients of any scalar targets w.r.t. intermediate features can be calculated automatically [67]. In this way, generating an ODAM explanation for one prediction takes about 2ms, which is much faster than the perturbation-based D-RISE, where using 5000 masks requires ∼3 minutes to process one image with FCOS.

4 Experiments with ODAM

In this section we conduct experiments on ODAM to: 1) evaluate its visual explanations qualitatively (Sec. 4.1) and quantitatively (Sec. 4.2), and compare with the current state-of-the-art methods; 2) evaluate the object specification and discrimination ability through user trust experiments, and also investigate how well the visual explanations of object detectors agrees with human explanations, as measured through eye gaze, and whether this agreement is related with user trust of the detector (Sec. 4.3). We mainly conduct the experiments with one-stage detector FCOS [18] and two-stage detector Faster R-CNN, [22] using ResNet-50 [68] as the backbone and FPN [69] as the neck. Two datasets are adopted for evaluation: MS COCO [70], a standard object detection dataset, and CrowdHuman [71], containing scenes with heavily overlapped objects. Experiments are performed using PyTorch and an RTX 3090 GPU. The training and testing hyperparameters are the same as those of the baseline detectors.

4.1 Qualitative evaluation of visual explanations

4.1.1 Comparison of different explanation methods

We compare the visualizations of Grad-CAM, Grad-CAM++, D-RISE and our ODAM in Fig. 1. The results are all generated with FCOS using FPN features on both the *MS COCO val* and *CrowdHuman val* sets. Grad-CAM, Grad-CAM++ and ODAM use class score targets, while D-RISE uses the same mask settings as [17] to find the attention area of predictions. Our ODAM demonstrates a strong ability of generating clear and distinct instance-specific heat maps, and generalizes well on normal/crowded scenes, and multiple-/single-class datasets. In contrast, Grad-CAM and Grad-CAM++ are class-specific (highlighting multiple objects with the same target class), and D-RISE contains "speckles" in the background due to its random masking mechanism. Comparisons with BBAM [41] are presented in the supplemental.

4.1.2 Different prediction attributes and detectors.

To verify the interpretability of visualizations, we generate explanations for different prediction attributes (class score, bbox regression values) of two specific instances using various detector architectures with the two types of backbones. To obtain a holistic view of the explanations generated by different models, we compute a combined heat map based on element-wise maximum of heat maps for the predicted class and bbox regression, $H_{\text{comb}} = \max(H_{\text{class}}, H_{x_1}, H_{y_1}, H_{x_2}, H_{y_2})$. Example results are presented in Fig. 3 (left), and we have the following observations from examining many such examples: 1) When predicting the object class, the model attends to the central areas of the object; 2) when regressing the bbox, the model focuses on the extent of the object, 3) For the same target, models from different detectors show attention on different regions, even though they all detect the same instance with a high confidence. Thus, developers can have an intuitive judgment about the model through explanation maps.

4.1.3 Different feature map layers.

In Fig. 4, we visualize and compare the heat maps generated from different feature layers. The feature maps A_k are selected from the output feature maps of the ResNet backbone (resnet_p3-p5), which are also the inputs of Feature Pyramid Network (FPN) with P3-P5 level, the FPN ouput, or the RoI pooling output. Note that the heat map for the RoI Pooling layer is obtained by bilinear interpolation and inserting the original map of 7×7 to the image plane. For the same target, higher-level layers (e.g. FPN and RoI pooling) show more concentrated attention and generate smoother heat maps. In contrast, for the lower-level layers, the heat

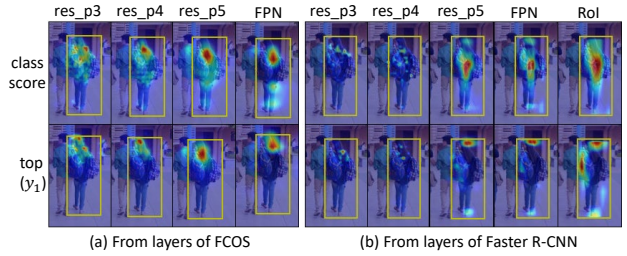


Fig. 4: Heat maps computed from different feature maps of one-stage FCOS and two-stage Faster R-CNN, when interpreting the class score and the regression of the top extent (y_1 in bounding box), respectively.

Fig. 5: Explanations of the predicted right extents (x_2 in bounding boxes) for different predictions of "surfboard". The heat maps for the mislocalized predictions highlight the visual features that induced the wrong predictions (e.g., the leg on the right, and the sea horizon).

maps show the individual feature locations that were important. Moreover, for this specific person in the image, when regressing y_1 (i.e., the top of the bbox), the head and upper body gets higher response than the legs, and thus we infer that both FCOS and Faster R-CNN concentrate mainly on the top extent of the object here. Note that the heat map of Faster R-CNN also highlights the person's feet, while FCOS does not. This reflects that the calculation of y_1 in Faster R-CNN is related to both the y-axis bias to the center and the ratio of the anchor height. In contrast, y_1 in FCOS is only related to the distance from the top to the center.

4.1.4 Analyzing error modes of the detector

Next we use ODAM to analyze the error modes of a detector. For the high confidence but poorly localized cases, we generate explanations of the wrong predicted extents and compare them with the correct localization results. As shown in Fig. 5, the model highlights that the wrong extents were misled by the leg of the person and the sea horizon.

To analyze classification decisions of the model, we generate explanations of the class scores. In Fig. 6(a), the model correctly classifies the "bed" object when seeing the cushion of the bed, but also mistakenly predicts "bench" based on a long metal bed frame at the end of the bed. In In Fig. 6(b), a person is fixing a wheel on the ground, and two motorcycles are parked nearby. The detector correctly finds the person, but also mistakenly detects a motorcycle on top of the person, by combining the features from the two motorcycles. This shows a failure mode of the detector, where sometimes the context feature (a person next to unrelated motorcycle parts) may negatively influence the detection result.

4.1.5 Comparison with attention maps of DETR.

The self-attention of the transformer architecture is a bottom-up attention process. Thus, visualizing the self-attention of transformer-based detectors, e.g. DETR, can show what features the detector was focusing on. In Fig. 7, we visualize the heat map explanations for DETR using ODAM, which is a *top-down* visual explanation, and the DETR transformer's self-attention, which is a *bottom-up* saliency map. For the encoder transformer, the self-attention will sometimes focus mainly on the object (e.g., 1st and 2nd rows of Fig. 7a), but also sometimes look at context features (e.g., in 3rd

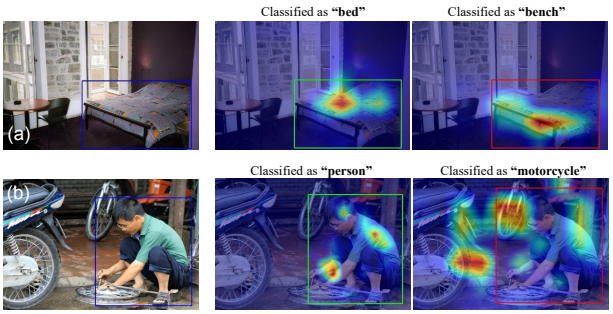


Fig. 6: Explanations of the class scores of different predictions. (a) The model predicts "bench" when it puts attention on only the frame at the end of the bed. (b) The model is negatively influenced by the context features and misclassifies a "motorcycle" on a "person".

and 4th rows of Fig. 7a, the bed surrounding the cat and the remote control). For the decoder transformer shown in Fig. 7b, the self-attention will look at the extremities of the object, i.e, the points along the predicted bounding box.

The ODAM heat maps for individual attributes are shown in Fig. 7(d) with the combo maps shown in the final column. The ODAM heat map for the *class score* is mostly consistent with the encoder self-attention maps. However in some cases (e.g., the remote control in the 4th row), ODAM shows that less context information is actually used compared to what is indicated by the encoder self-attention map. The ODAM combo maps highlight information consistent with the Generic Attention-model Explainability (GAME) [72], which is a visual interpretation method designed specifically for attention layer based models. However, in contrast to GAME, ODAM is also able to highlight the important regions for predicting *each coordinate* of the bounding box, e.g., both the rear and the back legs of the zebra are important for predicting the right box coordinate x_2. In contrast, the decoder self-attention highlights all extremities of the zebra, so the self-attention itself cannot disentangle the important regions for individual outputs, like each coordinate of the bbox. From the self-attention and ODAM visualizations, we may hypothesize how DETR performs detection: *the DETR encoder mainly aims to discover the object classes in the image through aggregation of class-related features, while the DETR decoder mainly aims to generate the bounding box prediction from the discovered objects by focusing on the extremities of the objects.*

The transformer self-attention map is a bottom-up attention map, i.e., generally showing which features are interesting and correlated with the query. For example, in the transformer encoder, with the feature itself as query, heat maps highlight the attention w.r.t. each location on the feature map. In the decoder, with the query corresponding to each prediction, the attention map shows the regions that are highly correlated with the query. In contrast, the ODAM generates a top-down attention map, i.e, which features are important for the output prediction. It should be noted that for bottom-up attention, even if a feature is highlighted in the attention map, there is no guarantee that the feature is actually used in the subsequent output prediction. In contrast, for the top-down attention, by design, all highlighted features should be important for generating the prediction.

4.1.6 Applying ODAM on GLIP and CLIP

Recently, several works on multi-modal representations [73, 74, 75, 76] have been developed to learn representations at the interaction of computer vision and natural language processing.

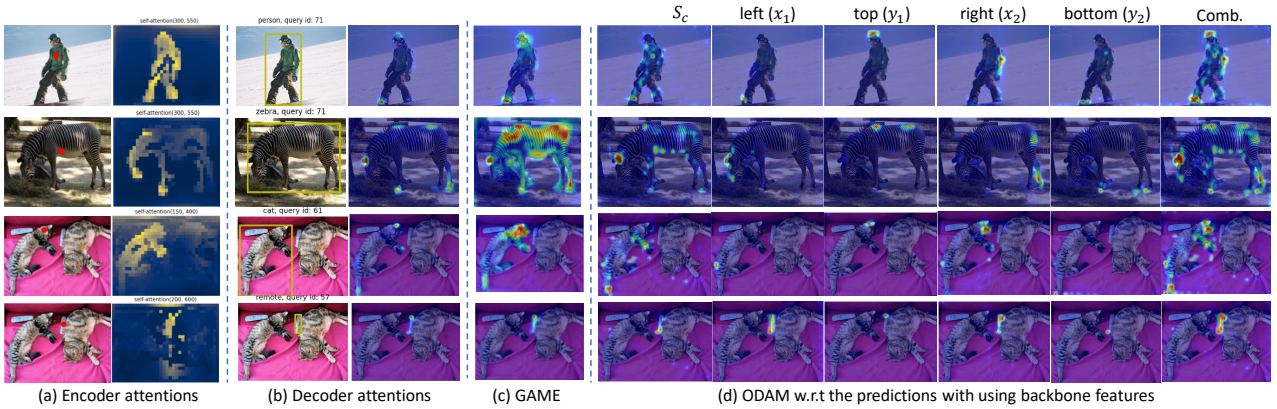


Fig. 7: Visualizations of self-attention maps in DETR and heat map explanations using ODAM. (a) The locations for querying the self-attention in the encoder, and the encoder self-attention weights at the positions specified; (b) The predictions of the detector, and the decoder self-attention weights of the predictions; (c) The interpretation generated by Generic Attention-model Explainability (GAME); (d) The ODAM heat maps of the prediction attributes using the backbone features, including maps for class score S_c, coordinates of the bounding box, and the combined (Comb.) map.

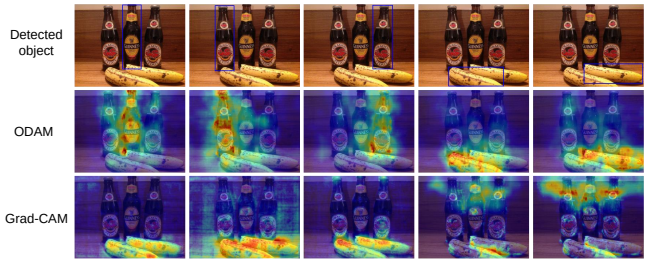


Fig. 8: Comparison of heat maps from our ODAM and Grad-CAM for interpreting predictions from the GLIP detector. Visual explanations are provided for predicted "bottle" and "banana" instances, and the blue boxes show the corresponding detected objects.

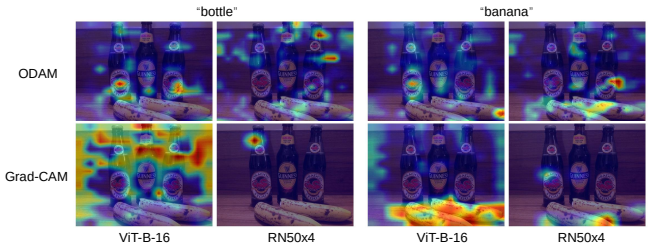


Fig. 9: Comparison of heat maps from our ODAM and Grad-CAM for interpreting image-text cosine similarity score from CLIP model.

The resulting pre-trained multi-modal models have good results on various down-stream tasks. Grounded Language-Image Pre-training (GLIP) [74] is a multi-modal model pre-training method for visual grounding, which has similar function as object detection. Thus, we implemented our ODAM on GLIP, and show explanation maps for the grounded results in Fig. 8. ODAM can generate explanation maps for each specific detected "bottle" and "banana", while the Grad-CAM fails to show a meaningful result. When aligning the region features and the corresponding object text prompts, GLIP also learns the region-aware representations besides the whole image feature. Therefore, ODAM is capable with GLIP to generate instance-specific explanation results.

For comparison, we also attempted to use ODAM and Grad-CAM to interpret the popular multi-modal pre-training model CLIP [73]. Specifically, we generate explanation maps on CLIP with the cosine similarity of the image and the corresponding text ("bottle" or "banana") as target. The gradient is calculated using the target w.r.t. the feature maps from ViT or ResNet50x4 layer. The results are shown in Fig. 9. Both the gradient-based methods Grad-CAM and ODAM perform poorly at explaining the multi-modal pre-trained model, with the possible reasons: (1) ODAM is sensitive to spatial information provided on the feature map, while CLIP learns the visual representation for the whole image and lacks of supervision for region-aware and localization ability; (2) For the whole image representation, special design may be required for explaining attention layer based aggregation, which is different from the CNN-based architectures.

4.2 Quantitative evaluation of visual explanations

We next perform quantitative evaluation of the ODAM visual explanations. We evaluate the ability of object specification (i.e, explanation faithfulness) using Deletion, Insertion [77, 11, 12, 13, 17] and visual explanation accuracy [78]. For evaluating localization, we adopt Pointing Games (PG) [35]. Meanwhile, we propose an object discrimination index (ODI) for measuring the interpretation ability of object discrimination.

For comparison, we implement D-RISE using 5000 masks with resolution of 16×16 as in [17]. FCOS is used as baseline model and the features from FPN are adopted to generate heat maps. The best matched predictions of the well-detected objects ($IoU > 0.9$) in the evaluation dataset are interpreted by each explanation method. The confidence scores of predictions are used as the explanation targets in ODAM, Grad-CAM and Grad-CAM++. Besides the *MS COCO val* set, results of PG and ODI are also reported on *CrowdHuman validation* sets.

4.2.1 Deletion and Insertion

A faithful heat map should highlight the important context on the image, which shows the ability of object specification. Deletion replaces input image pixels by random values step-by-step using the ordering of the heatmap (most important first), then measures the score drop of the predicted confidence (in percentage). Insertion is the reverse operation of Deletion, and adds image pixels to an empty image in each step (based on heatmap importance) and records the average score increase. Since the object sizes vary a lot, we consider each step as 1% of the bounding box area and record results in each step. For each step, each well-detected object is processed separately, and the scores are averaged.

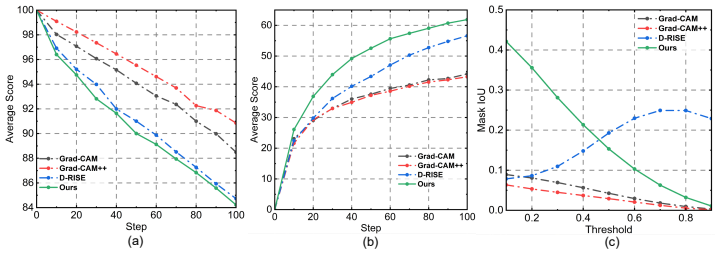


Fig. 10: Average prediction score vs. (a) Deletion steps and (b) Insertion steps. (c) The IoU between the ground truth object mask and thresholded explanation heat map.

TABLE 1: Evaluation of faithfulness on *MS COCO val* set: AUC for Deletion, Insertion and Visual Explanation Accuracy (VEA) curves in Fig. 10.

Method	Deletion↓	Insertion↑	VEA ↑
Grad-CAM	92.79	36.78	0.039
Grad-CAM++	92.52	36.18	0.027
D-RISE	73.35	43.35	0.157
ODAM	**72.68**	**50.33**	**0.163**

The average prediction score curves are presented in Fig. 10(a-b), and Tab. 1 reports the area under the curve (AUC). Lower Deletion AUC means steeper drops in score, while higher Insertion AUC means larger increase in score with each step. Our methods have the fastest performance drop and largest performance increase for Deletion and Insertion, which shows that the regions highlighted in our heat maps have larger effects on the detector predictions, as compared to other methods. Note that instance-specific explanations (ours and D-RISE) significantly surpass the explanation designed for image classification (Grad-CAM, Grad-CAM++), which are not instance-specific.

4.2.2 Visual explanation accuracy (VEA)

VEA measures the IoU between the ground-truth (GT) and the explanation heat map thresholded at different values. We use the MS COCO GT object masks for this evaluation, and results are presented in Fig. 10(c). Our method obtains the highest IoU when the threshold is small ($T < 0.4$), and the IoUs decrease as the threshold increases. This indicates that our heat map energy is almost all inside the object (consistent with energy PG results in Tab. 2). In contrast, the low IoUs of the previous methods at a small threshold indicate that their heatmaps contains significant amounts of energy outside the object mask. For Grad-CAM(++), this is because they are not instance-specific. In contrast, for D-RISE, the low initial IoU is caused by the mask sampling process. As the threshold increases, the IoU of D-RISE also increases, which indicates that the heat map values inside the object mask are larger than those outside the mask. Overall, our method has better IoU (0.421 at T=0.1) compared to D-RISE (0.249 at T=0.7), which is also confirmed by the VEA AUC in Tab. 1.

4.2.3 Pointing Game (PG)

We next quantitatively evaluate the localization ability of the ODAM explanations using the PG metric. To compute PG, the maximum point in the instance-level heat map is extracted and a hit is scored if the point lies within the GT object region (either bbox or instance mask). Then the PG accuracy is measured by averaging over the test objects. Since PG only considers the maximum point, but not the spread of the heat map, we also adopt the energy-based PG [12], which calculates the proportion of heat map energy within the GT object bbox or mask (versus the whole map). Finally, to show the *compactness* of the heat map, we calculate the weighted standard deviation of heat-map pixels, relative to the maximum point: $Comp. = \left(\frac{1}{\sum_x S_x}\sum_x \frac{S_x||x-\hat{x}||^2}{\frac{1}{4}(h^2+w^2)}\right)^{\frac{1}{2}}$, where S_x is the heat map value at location x, $\hat{x}$ is the maximum point of the heat map, and (w, h) are the width and height of the GT box. The denominator normalizes the distance w.r.t. the object size. Smaller compactness values indicate the heat map is more concentrated around its maximum.

The PG results are presented in Tab. 2. Grad-CAM and Grad-CAM++ perform poorly since they do not generate instance-specific heat maps. ODAM yields significant improvements over D-RISE on all the metrics. Specifically, D-RISE cannot work well on CrowdHuman, which only contains one object category. D-RISE uses the similarities between predictions of masked images and the original image to decide the mask weights, but for datasets with few object categories, the predicted class probabilities provide less useful information when calculating the weights.

4.2.4 Object Discrimination

To evaluate the object discrimination ability of the heat maps, we propose the *object discrimination index* (ODI), which measures the amount of heat map energy that leaks to other objects. Specifically, for a given target object, the ODI is the proportion of heat map energy inside all other objects' bboxes (regardless of class) w.r.t. the total heat map energy inside all objects in the image (i.e., ignoring background regions). Lower ODI indicates less heat map energy on other objects, and therefore better ability to discriminatively explain the detected object. ODI can also be computed using segmentation masks instead of bounding boxes.

The average ODIs are presented in Tab. 3. ODAM consistently shows the least energy leaking out to other objects, i.e., better object discrimination ability. Note that when using the tighter GT mask on MSCOCO, ODAM obtains the largest proportion of decrease, which indicates that the heat map can better focus on the explained target, even if its bbox overlaps with other objects.

4.3 User trust and human attention studies

In this section, we conduct user trust studies to investigate how well the interpretability of ODAM explanation maps can induce user trust in detectors. We also conduct a study to compare the visual explanations of the detector with human attention, and see whether it is correlated with user trust.

4.3.1 User trust study on object specification

For object specification, the interpretability of visual explanations is evaluated through the following user trust test. Heat maps are generated by D-RISE, Grad-CAM, Grad-CAM++, and ODAM for 120 correctly-detected objects out of 80 classes from the MSCOCO val set (1-2 instances for each class). For each object, users are asked to rank the maps from the four methods by the order of "which map gives more reasonable insight about how the target object was detected". We collect 10 responses for each object from a total number of 40 users (30 objects per user), totaling 1200 responses. The form of the questionnare and some samples are shown in Fig. 11 (left).

The results are presented in Tab. 4. ODAM is ranked 1st place in 53.8% of trials and 2nd place in 35.4% of trials, which is significantly better than D-RISE (χ^2 test, p<0.001). Overall, ODAM

TABLE 2: Comparison of Pointing Game (PG) accuracy with ground-truth bounding boxes or segmentation masks, energy-based PG (en-PG) with box or mask, and Heat Map Compactness (Comp.).

	MS COCO					CrowdHuman		
	PG(box)↑	PG(mask)↑	enPG(box) ↑	enPG(mask) ↑	Comp.↓	PG(box)↑	enPG(box) ↑	Comp.↓
Grad-CAM	26.7	22.5	20.7	15.0	4.34	15.7	9.7	3.99
Grad-CAM++	26.6	20.2	20.0	14.8	4.91	15.4	11.4	3.84
D-RISE	82.6	68.0	17.4	12.0	5.17	1.5	1.7	3.53
Ours	**91.9**	**82.6**	**73.1**	**57.1**	**1.36**	**95.5**	**79.5**	**1.04**

TABLE 4: User trust on object specification: (left) percentage of rankings for each method; (right) average rank (AR).

Method	1st	2nd	3rd	4th	AR
Grad-CAM	3.9	12.9	30.5	52.7	3.3
Grad-CAM++	7.3	22.2	43.1	27.5	2.9
D-RISE	35.1	29.5	17.5	17.9	2.2
ODAM	**53.8**	35.4	8.9	1.9	**1.6**

TABLE 5: User trust on object discrimination: (left) the percentage of each confidence level of user responses; (right) the average confidence, and the accuracy for correct object discrimination.

Method	1 (least)	2	3	4	5 (most)	avg. conf.	accuracy
Grad-CAM	53.38	30.41	10.14	5.41	0.68	1.70	14.19
Grad-CAM++	63.76	24.16	6.71	4.70	0.67	1.54	18.79
D-RISE	20.67	36.67	26.01	11.98	4.68	2.43	60.67
ODAM	0	0.67	7.33	21.34	70.68	**4.62**	**94.00**

TABLE 3: Comparison of Object Discrimination Index (%) using GT bbox or mask.

	MS COCO		CrowdHuman
	box ↓	mask ↓	box ↓
Grad-CAM	77.0	72.7	91.4
Grad-CAM++	77.3	73.2	92.0
D-RISE	71.0	66.3	95.3
ODAM	**34.8**	**19.5**	**56.9**

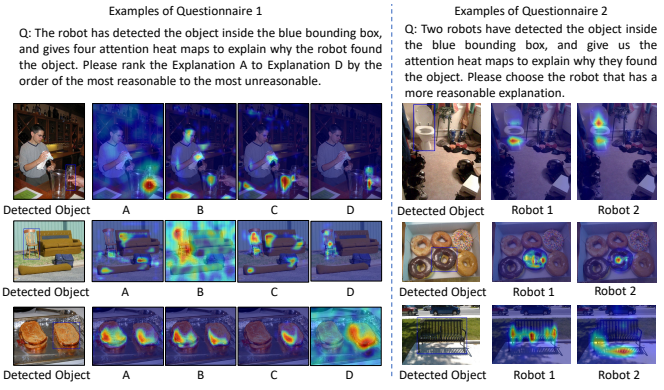


Fig. 11: User trust studies for object specification: (left) In the first questionnaire, the user needs to rank the heat maps from four methods: D-RISE, Grad-CAM, Grad-CAM++ and ODAM (ours). (right) In the second questionnaire, the user is asked to choose the more reasonable heat map from the two explanations, which are generated from detectors with different performance (36.6% mAP and 42.3% mAP). The labels for each option are assigned randomly for each question.

has significantly better average rank of 1.6 compared to other methods (Wilcoxon signed-rank test, $p<0.001$). The significantly higher human trust of ODAM demonstrate its superior instance-level interpretability for object detection.

From another aspect, previous studies evaluated trust between humans and DNNs by seeing if better models had better explanation maps according to humans. Following [5, 17], ODAM maps are generated for 120 objects that are correctly detected by two FCOS-ResNet50 models with different performance (36.6% mAP and 42.3% mAP). Some samples of this questionnaire are shown in Fig. 11 (right). Excluding samples where users thought the explanations were similar quality, the better model (42.3% mAP) received significantly more responses that its explanations were more trustworthy (38.2% vs. 28.6%; $\chi^2(1)$=8.10, p=0.004). Thus the ODAM visualizations provide evidence that more accurate models can induce more trust in humans.

4.3.2 *User trust study on object discrimination*

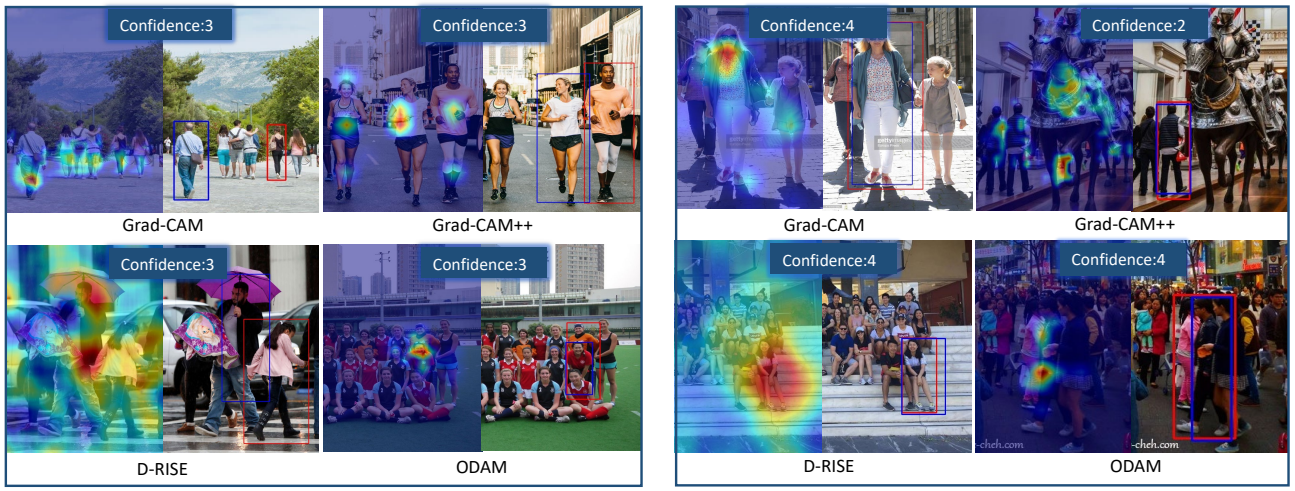
The previous section conducts the user study to evaluate the explanations for faithfulness, which shows how reasonable the heat maps explain the object class predictions. Here, we further conduct a user study on object discrimination ability. In this test, users are asked to draw a bounding box on the image to annotate which object they think the AI has detected, based on the shown heat map. Meanwhile, the users also need to provide their confidence level from least certain (1) to most certain (5) when making the choice. For each method (D-RISE, Grad-CAM, Grad-CAM++ and ODAM), 150 samples are sent out to 10 users with one user annotating on 15 images. Since the purpose is to test whether the heat maps can effectively show which object was detected, especially in the crowded scene, we choose the samples from the CrowdHuman validation set. Since users will have different ways to draw the box around the object, a separate marker manually inspected the user's boxes to see if they align with the GT object box in order to determine correctness. During this process, the marker does not know which explanation method was used for each image.

Tab. 5 presents the number of examples (in percentage) under each confidence level and the accuracy of users' decisions (the ratio of users' correct choices) based on heat maps of each method. The results show that the user can obtain a much higher accuracy (χ^2 test, $p<.0001$) and higher confidence (t-test, $p<.0001$) with heat maps from ODAM, which demonstrates that ODAM explanations are superior on object discrimination ability, inducing higher user trust. These results are consistent with the quantitative evalution with ODI in Sec. 4.2. Some incorrect and correct user's choices are displayed in Fig. 12.

4.3.3 *Similarity with human attention*

Since ODAM achieves the highest human trust in the above user studies, we further investigate how well the visual explanations of object detectors agree with human attention, and whether this agreement is related with user trust of the detector. The human attention data is collected using an explanation task, where the user describes a specific object in an image using a text explanation, and meanwhile their eye fixations are recorded using a hardware eye tracker[1]. Here we use an explanation task, where the participant must describe the discriminative features of the object,

1. The experiment has been approved by the Human Research Ethics Committee at the University of Hong Kong (Reference number : EA210386)



(a) Examples of user's incorrect choice (b) Examples of user's correct choice

Fig. 12: User trust studies for object discrimination. Examples of (a) user's incorrect choices and (b) user's correct choices with heat maps of Grad-CAM, Grad-CAM++, D-RISE and ODAM, respectively. In the user trust test on object discrimination, users are asked to draw the bounding box of the object which was detected based on the given heat map. Blue boxes are those drawn by users, while red boxes are those of the ground truth objects. Note that in CrowdHuman, the ground-truth boxes are for the full person, even when the person is partially occluded. The user's confidences when choosing the objects are also displayed.

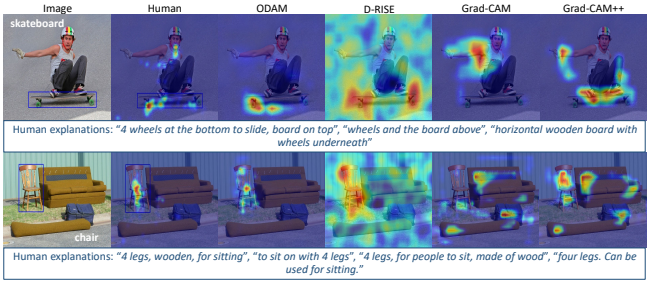


Fig. 13: Examples of human attention maps and corresponding text explanations, and XAI visual explanation heat maps for the FCOS detector using ODAM, D-RISE, Grad-CAM and Grad-CAM++.

TABLE 6: Quantitative comparison of human attention maps and XAI visual explanations for FCOS using: cosine similarity (Cos.), correlation (Corr.), optimal transport distance (OTD), and Jensen-Shannon divergence (JSD).

Method	Cos.↑	Corr. ($\times 10^{-3}$)↑	OTD↓	JSD↓
Grad-CAM	0.1802	0.467	0.2483	0.5216
Grad-CAM++	0.2242	0.674	0.2225	0.5051
D-RISE	0.4142	0.438	0.1315	0.3902
ODAM	**0.4627**	**2.876**	**0.0653**	**0.3180**

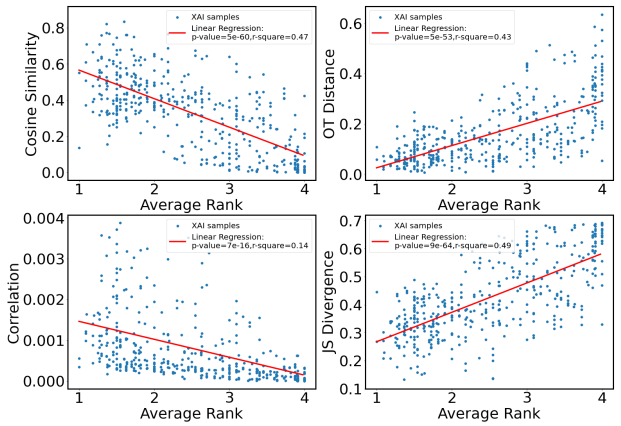


Fig. 14: Relationship between user trust of an ODAM visual explanation maps and its similarity to human attention maps. The user trust is the average rank of the ODAM maps obtained in the user study in Fig. 11 left. The similarity/distance between the visual explanation the human attention map is measured with cosine similarity, OT (optimal transport) distance, correlation and JS (Jensen-Shannon) divergence between ODAM maps and human attention maps.

rather than a recognition task, where the participant just names the object, because previous studies [79, 80] have shown that human attention during explanation is more similar to XAI heat maps for image classification. In contrast to [14, 81] that use a manual pointing/reveal task to collect the human attention, collecting eye fixations is a more direct measurement of top-down attention while the user performs explanation tasks.

In detail, two example objects with high detection confidence were selected from each class in the *MS COCO val* set, totaling 160 objects. We collected eye fixation data during an explanation task of the query object's class (80 classes) from 10 participants. All participants had normal or corrected-to-normal vision. In each trial, the participant viewed an image with the query object annotated by a bounding box. All the images are displayed on a 1280x1024 screen, and images are scaled so that all query objects are similar sized and large enough to discriminate eye fixations on object parts. Specifically, each image is resized and padded such that the query object has the minimum side length of 10° of visual angle (v.a.) or ∼230 pixels. Meanwhile, the object area is limited to 440^2 pixels to prevent some stick-shaped objects from becoming too big. The participant was then asked to explain why this object is a particular class, and at the same time, the eye fixation locations and durations are recorded using an EyeLink 1000 hardware eye tracker. For each image, the eye fixations of the ten participants were combined, and an overall human attention map was computed by convolving the eye fixation map with a 2D Gaussian kernel with bandwidth 23 pixels (equivalent to 1° v.a.). Finally the eye gaze maps are transformed back to the original image size for visualization and comparison with XAI heat maps.

Fig. 13 presents two examples of the human attention maps and XAI heat maps generated for the FCOS detector. The similar-

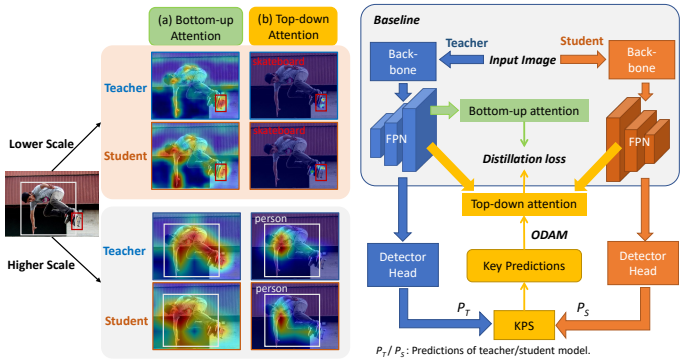


Fig. 15: ODAM-based knowledge distillation (KD). Left: An example comparing (a) bottom-up attention and (b) top-down attention (ODAM explanation map) of the teacher model (RetinaNet-Resnet101) and student model (RetinaNet-Resnet50). The top-down attention is generated using ODAM w.r.t the key prediction in the lower scale (skateboard) and the higher scale (person), respectively. Right: our proposed framework of KD using ODAM explanation maps.

ity between the human attention maps and the XAI heat maps is computed using 4 similarities/distances: cosine similarity, correlation, optimal transport distance (OTD), and Jensen-Shannon divergence (JSD). The average results are shown in Tab. 6. Compared to other XAI explanation maps, ODAM achieves higher similarity (Cosine, Correlation) with and smaller distance (OTD, JSD) to human attention maps. Since ODAM obtains higher confidence in the user trust study as well, the consistent results give a promising insight on the relationship between user trust of explainable AI and the similarity with human attention. We further investigate this relationship by plotting the similarity (distance) with human attention vs. the average rank in the user trust study for each visual explanation map. The scatter plots in Fig. 14 also show a statistically-significant trend (red line) that visual explanations that induce higher confidence (better ranking) in the user trust study have higher similarity (smaller distance) with human attention.

5 Application of Object Specification: ODAM-based Knowledge Distillation

For the explanation task of "object specification", the heat map highlights the important regions for the specific prediction, and providing an interpretation about "what context/features are important for the prediction?". ODAM is able to generate instance-level explainable maps, which shows the top-down attention of the detector for a specific prediction. Therefore, we propose an application for the object specification ability of ODAM: ODAM-based knowledge distillation (KD) for object detection. With the top-down attention provided by ODAM, a student detector is expected to learn better from the teacher detector.

5.1 Brief introduction of KD for object detection

Larger backbones are usually needed to obtain higher-accuracy from deep learning based models, which thus consume more computing resources and inference time. Knowledge distillation (KD) [47] was proposed to transfer the learned information from a large teacher model to a lightweight student network. By mimicking the predictions or features of the teacher, the student can achieve higher performance with lower memory requirements and inference time. The KD methods [82, 48, 83, 84, 85, 86, 87, 88] designed for image classification show less effectiveness when directly migrated to an object detection model, since there are multiple proposals comprising an extremely unbalanced ratio of positive and negative instances in the detection task. Therefore, for KD in object detection, a key issue is to select the imitation region and decide the importance in features or predictions.

The state-of-the-art FGD [60] distills object regions and background regions separately, with using the bottom-up attention map as the weight of feature distillation. Since a Feature Pyramid Network (FPN) [69] is typically used in detectors to extract features of different scales, KD generally occurs on each level of FPN. For each level, the teacher's feature map is utilized to generate the bottom-up attention as: $M_{bu} = C \cdot \text{softmax}(\frac{1}{K}\sum_k |A_k^{(T)}|)$, where $A_k^{(T)}$ is the feature map of the k-th channel of the teacher network, and C is the spatial area (number of pixels) of the feature map. Since the foreground and background are distilled separately, a normalization mask N is defined to normalize separately with respect to the object or background area: $N_{ij} = \frac{1}{C_{obj}}$ if location (i, j) is in an object, or $N_{ij} = \frac{1}{C_{bg}}$ otherwise, where C_{obj} and C_{bg} are the areas of the object and the background respectively. Finally, the feature distillation loss is a weighted distance of the features between teacher and student: $Loss = \sum_{ij} N \circ M_{bu} \circ \sum_k L_2(A^{(T)}, A^{(S)})$, where L_2 is the pixel-wise L2 between feature vectors, the summation of k is over channels, the summation over ij is over space, and $\circ$ is the element-wise product.

5.2 ODAM-based knowledge distillation

We next propose using ODAM to generate top-down attention for knowledge distillation. In order to explore the important features for predictions on each scale and the differences between students and teachers, we visualized the bottom-up attention M_{bu} and the top-down attention generated from ODAM in Fig. 15(left). The bottom-up attention (Fig. 15 left(a)) shows the magnitude of feature values on each location where higher values indicate larger feature responses. The top-down attentions (Fig. 15 left(b)) are generated with ODAM and show which features are actually important for the detector to make a specific prediction. For different feature scales, when predicting objects in various sizes, not all foreground features contribute to the final predictions (e.g. the small object "skateboard" predicted with lower scale feature is not used to predict the large object "person" with higher scale feature). Therefore, we propose an ODAM-based KD that first selects the key predictions in each feature level, and then uses ODAM to indicate the focused region of the foreground. The focused region and unfocused region may lead to different difficulties and effectiveness in knowledge distillation.

The framework of our ODAM-KD is shown in Fig. 15 (right), which uses FGD as the baseline. The key prediction selection (KPS) module selects the key outputs of the teacher and student models, which are passed as input to ODAM to generate the top-down attention map. We consider two aspects in the selection, teacher's strengths and student's weakness, which should influence the performance gap between student and teacher, so that a teacher can make effective and precise knowledge transmission.

For the teacher model, after the anchor assignment process[2] the top 10 best matched predictions for each ground truth object (GT) are selected and checked, then those whose IoU with their GT box is larger than 0.5 are regarded as the key instances. For the student model, we regard the current false positive predictions

2. During detector training, anchor assignment will choose positive predictions for each ground truth object based on the matching quality.

as key instances, since they could cause negative influence to the student's performance. Therefore, the student predictions are selected to satisfy high classification confidence (over 0.1), which makes it survive at inference, and bad IoU with its assigned GT (below 0.5), which means a mislocalization. We record the corresponding anchor locations of both the selected teacher's high-quality and student's low-quality predictions in the set $\mathcal{P}$, and denote the object confidence outputs for prediction p as $y_p^{(T)}$ and $y_p^{(S)}$ for the teacher model and student model, respectively. Their corresponding GT objects are marked as **focused objects** in the specific feature level.

The gradients of predictions in $\mathcal{P}$ w.r.t. the feature map are calculated, and their weighted average is computed. Here the weights are the absolute difference of y_p in order to focus on predictions with large mismatch between teacher and student. Based on the weighted average gradients of teacher and student, we obtain attention maps $M_{td}^{(T)}$ with the features of teacher $A^{(T)}$, and $M_{td}^{(S)}$ with the student's feature $A^{(S)}$ in the form of ODAM:

$$M_{td}^{(T)} = \mathrm{ReLU}\Big(\sum_k A_k^{(T)} \circ \sum_{p\in\mathcal{P}} \Big|y_p^{(T)} - y_p^{(S)}\Big| \frac{\partial y_p^{(T)}}{\partial A_k^{(T)}}\Big), \quad (3)$$

$$M_{td}^{(S)} = \mathrm{ReLU}\Big(\sum_k A_k^{(S)} \circ \sum_{p\in\mathcal{P}} \Big|y_p^{(T)} - y_p^{(S)}\Big| \frac{\partial y_p^{(S)}}{\partial A_k^{(S)}}\Big), \quad (4)$$

$$M_{td} = C \cdot \mathrm{softmax}(\max(M_{td}^{(T)}, M_{td}^{(S)})), \quad (5)$$

where the combined map M_{td} is obtained with a softmax function is over the whole map.

Finally, with the bottom-up and top-down ODAM attention, we define the loss attention mask and feature distillation loss,

$$M_{att} = M_{bu} \circ M_{td}, \quad (6)$$

$$Loss = \sum_{ij} N \circ M_{att} \circ \sum_k L_2(A^{(T)}, A^{(S)}), \quad (7)$$

where N is the normalization mask defined in Sec. 5.1, while using the focused objects (not all GT objects) in its calculation.

5.3 Experiments with ODAM-KD

In these experiments we use ODAM explanation maps for knowledge distillation of detectors.

5.3.1 Ablation study

We first conduct an ablation study on ODAM-KD. The baseline distills all objects region as foreground on each level equally, while in contrast our method selects the *focused* GTs via KPS and only regards the locations inside their boxes as foreground. In the ablation study, we explore the contribution of focused and other GTs region to distillation. Here the other objects outside the KPS selection results are denoted as *unfocused* GT objects. After KPS selects the focused GT objects on each feature level based on the P^T and P^S, we conduct the experiments that use focused and unfocused GTs regions separately or together as the foreground with $N_{ij} = \frac{1}{C_{obj}}$ if location (i, j) is in foreground, otherwise $N_{ij} = 0$ for background, with C_{obj} denotes the object area.

As shown in Tab. 7, using the standard bottom-up attention, distillation on the features of focused regions brings the main effect to the performance (row 2 vs. row 1; mAP 39.5 focused vs. 39.1 unfocused), while distilling both parts together equally brings no improvement. Using our proposed attention map, which incorporates top-down ODAM attention to stress the important locations of the selected key instances (focused), the performance is further improved (row 4). These results demonstrate that the designed KPS module selects the discriminative and representative instances and corresponding GTs on each feature level, and the ODAM heat map of key instances helps to generate better attention for feature distillation. When also involving distillation on the background (row 6), our method further improves the performance (mAP 40.1 vs. 39.7 baseline, mAR 56.5 vs. 56.1 baseline).

TABLE 7: Ablation study for ODAM-KD: comparison of knowledge distillation in cases of: using focused, unfocused GTs region separately or together as foreground (FG); with or without distillation on background (BG); replacing the top-down attention with Grad-CAM.

RetinaNet [24]	FG	BG	Attention	mAP	mAR
ResNet101(T)	-	-	-	38.9	54.8
ResNet50(S)	-	-	-	37.4	53.9
Distillation (ResNet50)	(1) unfocused		bottom-up	39.1	55.3
	(2) focused		bottom-up	39.5	56.1
	(3) together (baseline)		bottom-up	39.5	55.9
	(4) focused (ours)		Eq. 6 w. ODAM	**39.8**	**56.3**
	(5) together (baseline)	√	bottom-up	39.7	56.1
	(6) focused (ours)	√	Eq. 6 w. ODAM	**40.1**	**56.5**
	(7) focused	√	Eq. 6 w. Grad-CAM	39.5	56.2

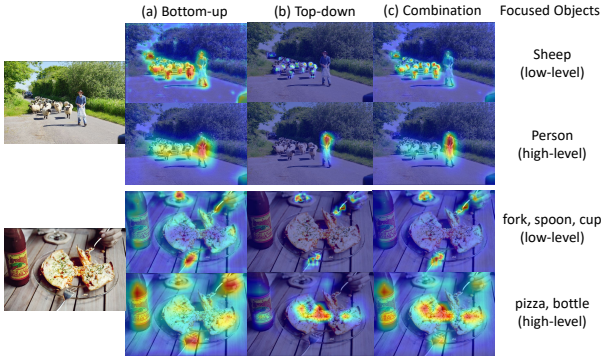


Fig. 16: Examples of bottom-up M_{bu} and top-down attention M_{td} and their combination M_{att} for different feature levels (spatial scales).

We further conduct an ablation study that replaces the ODAM top-down attention map used in (6) with Grad-CAM. Note that D-RISE is infeasible here since it is too inefficient to be calculated in every batch. The mAP using Grad-CAM (row 7) is lower than the baseline, which demonstrates that the regions highlighted by Grad-CAM are less appropriate for distillation. This also confirms that ODAM has superior object specification ability on detection.

Fig. 16 visualizes the bottom-up (M_{bu}), top-down (M_{td}), and combined (M_{att}) attention maps for 2 example images. The locations of foreground objects show higher magnitude of values on bottom-up attention (Fig. 16a), while not all the significant features have effects on the final predictions as seen in the corresponding top-down attention (Fig. 16b). For different feature levels, the KPS module is utilized to find the focused GT objects and key predictions, which are the inputs to ODAM to generate the top-down attention (Fig. 16b). By applying the "object specification" ability of ODAM, the top-down attention indicates the regions that are important for the model to replicate the selected representative predictions. Finally, Fig. 16c shows the combined attention maps, which consider both the significant feature areas and stresses the important areas for the key predictions.

5.3.2 Comparison with baseline method

In the experiment, FGD [60] is adopted as the baseline. We modify the attention map used in feature distillation of FGD to be generated by the designed KPS and ODAM. All other modules, including global distillation and training settings, follow the baseline. We implement the knowledge distillation on FCOS

TABLE 8: Comparison of knowledge distillation methods on MS COCO validation set.

Teacher	Student	Year	mAP	AP_S	AP_M	AP_L
RetinaNet-Res101 (mAP 38.9)	RetinaNet-Res50	-	37.4	20.6	40.7	49.7
	FGFI [56]	CVPR2019	38.6	21.4	42.5	51.5
	GID [55]	CVPR2021	39.1	**22.8**	43.1	52.3
	FGD [60]	CVPR2022	39.7	22.0	43.7	53.6
	Ours	-	**40.1**	**22.8**	**44.0**	**54.1**
RetinaNet-ResNeXt101 (mAP 41.0)	RetinaNet-Res50	-	37.4	20.6	40.7	49.7
	FKD [59]	ICLR2020	39.6	22.7	43.3	52.5
	CWD [89]	ICCV2021	40.8	22.7	44.5	**55.3**
	FGD [60]	CVPR2022	40.7	22.9	45.0	54.7
	Ours	-	**41.0**	**24.0**	**45.3**	55.0
FCOS-Res101 (mAP 40.8)	FCOS-Res50	-	38.5	21.9	42.8	48.6
	GID [55]	CVPR2021	42.0	25.6	45.8	54.3
	FGD [60]	CVPR2022	42.7	27.2	46.5	**55.5**
	Ours	-	**42.8**	**27.4**	**46.7**	55.0

[18] and RetinaNet [25] with MS COCO dataset [70]. The student model uses ResNet50 as backbone, and learn from the teacher models which uses a larger backbone, ResNet101 or ResNeXt101.

Tab. 8 presents the results for KD. Comparing with the FGD, ODAM-KD obtained consistently higher mAP with three teacher-student combinations, but with sometimes limited improvement (0.4 with RetinaNet-Res101, 0.3 with RetinaNet-ResNeXt101 and 0.1 with FCOS-Res101). Note that the gains on small objects are significant; AP_S improves by 0.8 with RetinaNet-Res101, 1.1 with RetinaNet-ResNeXt101, and 0.2 with FCOS-Res101. The reason why ODAM-KD improves performance on small objects is because the ODAM top-down attention map is more focused than the bottom-up attention map on small objects. Thus, their combination map can highlight each specific small object. This is illustrated in the visualization results in Fig. 16, where the combination attention has better focus on individual small objects, as compared to the bottom-up attention. Since ODAM-KD uses the same framework and loss design as FGD, with the difference being the distillation weights based on the combination attention, it is reasonable that ODAM-KD performance is significantly improved on AP_S, while obtaining slight gain on mAP. Overall, our method outperforms the baseline FGD for a variety of teacher backbones and detector architectures, and obtains the highest mAP among the recent object detection KD methods. The results further confirm the good object specification ability of ODAM explanation heat map, which provides effective attention on the important regions for KD.

6 Application of Object Discrimination: ODAM-based NMS

We now consider the unique explanation task for object detection, "object discrimination". This visual explanation map is expected to show which instance was looked at when the model made the prediction, and make interpretation about "which object was actually detected?" Here, we utilize the object discrimination ability of ODAM map and propose the ODAM-NMS to aid with duplicate removal while preserving overlapping detections of different instances in crowded scenes, where more predictions are likely to be mistakenly suppressed using classic NMS. In §6.1, we first introduce a training scheme, ODAM-Train, which encourages the model to generate heat map with better object discrimination ability. Then, in §6.2, we introduce our ODAM-NMS.

6.1 ODAM-Train

Since the instance-specific heat maps may still "leak" onto other neighboring objects, especially in crowded scenes, we first propose a training method ODAM-Train for improving the heat maps for object discrimination, to better explain which object was being detected. In order to focus the detector to be better localized on a specific object area, and not overlapped with other objects, ODAM-Train encourages similar attention for different predictions of the same object, and separate attentions for different objects (see Fig. 17a). Specifically, we propose a heat-map consistency loss L_{con} and a separation loss L_{sep} as auxiliary losses during detector training. Using the predicted confidence scores as explanation targets, heat maps are first calculated by ODAM for all the positive proposals, and then resized to the same size, and vectorized. The ODAM vectors are then organized by ground-truth (GT) object, where $\mathcal{P}^{(p)} = \{H_n^{(p)}\}_n$ is the set of ODAM vectors for positive predictions of the p-th GT object. For each GT object, the best prediction $H_{\text{best}}^{(p)}$ is selected from $\mathcal{P}^{(p)}$ that has the highest IoU with the GT box. The *consistency and separation losses* are:

$$L_{\text{con}} = \sum_{p \in GT} \sum_{n \in \mathcal{P}^{(p)}} -\log \cos(H_{\text{best}}^{(p)}, H_n^{(p)}), \quad (8)$$

$$L_{\text{sep}} = \sum_{p \in GT} \sum_{m \notin \mathcal{P}^{(p)}} -\log \left(1 - \cos(H_{\text{best}}^{(p)}, H_m^{(\neg p)})\right), \quad (9)$$

where $H_m^{(\neg p)}$ are ODAM vectors for proposals not corresponding to the n-th GT object. The loss for detector training is $L = L_{\text{detector}} + (L_{\text{con}} + L_{\text{sep}})/N$, where N is the total number of ODAM vector pairs in the loss calculation. Note that the proposed ODAM-Train is a fair design because the supervision of the heat map affects all layers of the detector via the top-down gradient computation of ODAM. Therefore, changes in the heat maps from learning (e.g., better localization on the object) are a result of *actual* changes in the detector's strategy (e.g., mainly using features on the object). Thus, in this way, the detector is trained to use a strategy that is more discriminative of object instances.

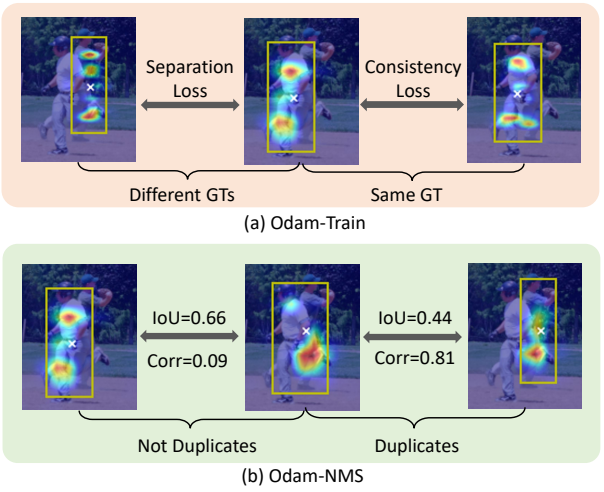


Fig. 17: (a) ODAM-Train uses auxiliary losses to encourage heat maps for predictions on the same object to be consistent, and for different objects to be distinct; (b) ODAM-NMS uses the box IoU and the normalized correlation between heat maps to determine the duplicate detections. The yellow bounding box shows the detection proposal corresponding to the heat map.

6.2 ODAM-NMS

In object detection, duplicated detections are removed in post-processing using NMS, which is based on the following assumption: two bounding boxes (bboxes) that are overlapped (high IoU) are likely to be duplicated, and the bbox with lower-score (less

Algorithm 1 ODAM-NMS: predictions are removed or kept based on both the IoU and the correlation between ODAM heat maps.

$P \leftarrow GetPredictions(imageI)$
$P \leftarrow SORT(P)$
$D \leftarrow \emptyset$
while $P \neq \emptyset$ **do**
  $p \leftarrow POP(P)$
  $isDuplicate \leftarrow false$
  for $d \in D$ **do**
    $iou \leftarrow GetIoU(p, d)$
    $corr \leftarrow NormCorrelation(S_{y_c}^{(p)}, S_{y_c}^{(d)})$
    if $iou \geq T_{iou}$ **and** $corr > T^l$ **then**
      $isDuplicate \leftarrow true$
    else if $iou < T_{iou}$ **and** $corr > T^h$ **then**
      $isDuplicate \leftarrow true$
    end if
  end for
  if $\neg isDuplicate$ **then**
    $PUSH(p, D)$
  end if
end while

confidence) should be removed. In particular, for classical NMS, the predictions in list P are sorted by their confidence scores, then each prediction $p \in P$ is compared to the currently selected detections $d \in D$. If the IoU between p and any $d \in D$ is larger than a threshold T_{iou}, p is considered a duplicate and discarded. Otherwise p is added to D. The classical NMS has shortcomings in crowded scenes because the key assumption does not hold when objects are partially occluded by neighboring objects.

We propose ODAM-NMS to mitigate this problem based on an observation that with ODAM-Train, the ODAM heat maps for different objects can be distinctive, even though their bboxes are heavily overlapped (see the left and center heat maps in Fig. 17b). Meanwhile, even if the IoU of two predicted bboxes is small, their visual explanations may be similar indicating that the same object instance is detected. For example, in Fig. 17b, the center and right predictions have low IoU but have heat maps with high correlation. In other words, ODAM shows which object the model looked at to make the prediction, which can intuitively assist NMS to better identify duplicate object predictions.

After the inference stage of the detector, we use ODAM to generate heat maps for each prediction with the predicted confidence score. All the heat maps are resized to the same size with a short edge length of 50, then vectorized. Normalized correlation is calculated between each pair of vectors to represent the probability that the two predictions correspond to the same object. ODAM-NMS uses both the IoUs and heat map correlations between p and $d \in D$ when considering whether a prediction should be removed or kept. If the IoU is large ($IoU \geq T_{iou}$) and the correlation is very small ($corr \leq T^l$), then p is not a duplicate; If the IoU is small ($IoU < T_{iou}$) and the correlation is very large ($corr > T^h$), then p is a duplicate. Through these two conditions, ODAM-NMS keeps more proposals in the high-IoU range for detecting highly-overlapped crowded objects, and removes more proposals in the low-IoU range for reducing duplicates. The pseudo code for ODAM-NMS is shown in Alg. 1.

6.3 Experiments with ODAM-NMS

We first present results on how ODAM-Train can improve the object discrimination ability of a detector. We then present results of ODAM-NMS using an ODAM-Trained detector on crowded scenes, followed by an ablation study comparing ODAM-NMS with and without ODAM-Train.

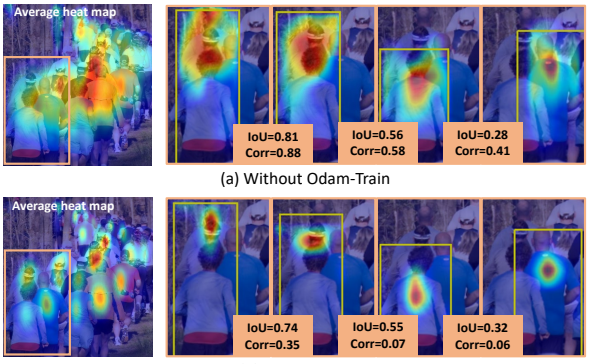


Fig. 18: Comparison of heat maps from FCOS without and with ODAM-Train. (left) The average heat map over the high-quality predictions with confidence score over 0.1. (right) Instance-specific heat maps of some predictions on different objects, with the IoU and correlation between each pair of predictions displayed in the middle.

6.3.1 The effect of ODAM-Train on object discrimination

We propose ODAM-Train to encourage the model to generate ODAM maps with better object discrimination ability. Here we compare the heat maps from ODAM using the baseline detector trained with and without ODAM-Train. The heat maps for each proposal are computed by ODAM with its confidence score as the explanation target. Although the original heat maps without ODAM-Train (Fig. 18a) can locate the object well, the attention may spread to its overlapping neighbors, which makes the correlations between them relatively high. Using the consistency and separation losses in (9), ODAM-Train yields well-localized heat maps for the same object and distinctive heat maps for different objects, which better shows which object was being detected, improving object discrimination. As seen in Fig. 18b, different people in the crowd can be separated by their low heat-map correlation, even when they have high bbox IoU.

We quantitatively compare the ODAM maps for detectors without and with ODAM-Train in terms of localization and object discrimination In Tab. 9, and in terms of object specification in Tab. 11. The higher PG, Compactness and ODI results indicate that using ODAM-Train can improve the localization quality and object discrimination ability by generating heat maps that are better-localized on the objects. However, there is a tradeoff with faithfulness (see Tab. 11), since the corresponding AUC metrics worsen slightly when using ODAM-Train.

6.3.2 Performance of ODAM-NMS with ODAM-Train

We evaluate ODAM-NMS with ODAM-Train for improving detection in crowded scenes. To evaluate the performance of NMS on heavily overlapped situations, we adopt the CrowdHuman dataset, which contains an average 22.6 objects per image (2.4 overlapped objects). We compare ODAM-NMS with NMS, SoftNMS, and FeatureNMS, using both FCOS and Faster RCNN. The IoU threshold is set to $T_{iou} = 0.5$ for both NMS and our method. Soft-NMS uses Gaussian decay with $\sigma = 0.5$ and final threshold of 0.05. For FeatureNMS, the IoU range is set to $(0.9, 0.1)$ following [66]. For ODAM-Train, the detector training pipeline is totally the same as the baseline [18, 22], which uses SGD as the optimizer with batch-size 16, learning rate 0.2 for two-stage Faster R-CNN

TABLE 9: Comparison of ODAM w/o and w/ ODAM-Train on the ability of localization and object discrimination, according to Pointing Game (PG) accuracy with ground-truth bounding boxes (b) or segmentation masks (m), energy-based PG with box or mask, Heat Map Compactness (Comp.), and Object Discrimination Index (ODI).

	MS COCO							CrowdHuman			
	PG(b)↑	PG(m)↑	enPG(b) ↑	enPG(m) ↑	Comp.↓	ODI(b)↓	ODI(m)↓	PG(b)↑	enPG(b) ↑	Comp.↓	ODI(b)↓
ODAM	91.9	82.6	73.1	57.1	1.36	34.8	19.5	95.5	79.5	1.04	56.9
w/ ODAM-Train	**93.3**	**83.9**	**79.6**	**63.9**	**1.32**	**34.1**	**18.7**	**97.3**	**83.9**	**0.91**	**51.3**

TABLE 10: Comparisons of NMS strategies on CrowdHuman validation set. All models are trained with the same baseline implementation. The timing is for the whole pipeline: detector inference, heat map calculation, and NMS.

	FCOS					Faster RCNN				
	AP↑	JI↑	MR↓	Recall	time (s/img)	AP↑	JI↑	MR↓	Recall	time (s/img)
NMS	87.8	78.4	45.5	93.2	0.114	86.9	79.5	43.2	90.3	0.092
Soft-NMS	80.8	74.9	89.0	93.0	0.470	76.5	61.9	84.8	92.3	0.284
FeatureNMS	89.3	78.1	45.6	95.4	0.145	82.0	65.7	68.8	**94.9**	0.120
ODAM-NMS (ours)	**89.3**	**81.1**	**44.5**	**95.5**	0.178	**88.1**	**80.5**	**42.8**	91.5	0.140

TABLE 11: Comparison of ODAM w/o and w/ ODAM-Train on object specification (explanation faithfulness) using AUC for Deletion, Insertion, and Visual explanation accuracy (VEA) curves.

	Deletion↓	Insertion↑	VEA ↑
ODAM	**72.68**	**50.33**	**0.163**
w/ ODAM-Train	74.45	46.66	0.143

TABLE 12: Comparisons of NMS performances with and without ODAM-Train (Od.-Tr.) on the CrowdHuman validation set for both FCOS and Faster RCNN detectors. T^h and T^l are the correlation thresholds in Alg. 1. The IoU threshold for all NMS methods is 0.5.

				FCOS			Faster RCNN		
	Od.-Tr.	T^h	T^l	AP↑	JI↑	MR↓	AP↑	JI↑	MR↓
NMS		-	-	87.8	78.4	45.5	86.9	79.5	43.2
	√	-	-	87.5	78.9	45.6	87.0	79.3	43.8
ODAM-NMS		0.8	0.2	87.5	78.5	54.0	88.9	78.7	44.3
		0.9	0.1	88.6	80.4	45.5	**89.0**	79.8	43.7
	√	0.8	0.2	**89.3**	**81.1**	**44.5**	88.1	**80.5**	**42.8**

and learning rate 0.1 for FCOS. The aspect ratios of the anchors in Faster R-CNN are set to $H : W = \{1, 2, 3\}$ based on the dataset, and other parameters are the same as in the baselines. Training runs for 30 epochs. We use $T^h = 0.8$, $T^l = 0.2$, which achieves stable performance in practice.

For comparisons, we adopt three evaluation criteria: *Average Precision* (AP_{50}); *Log-Average Missing Rate* (MR), which is sensitive to false positives (FPs), and commonly used in pedestrian detection; *Jaccard Index* (JI). See [90] for details. Smaller MR indicates better results, while larger AP_{50} and JI are better.

Tab. 10 shows the results. Soft-NMS performs poorly in crowd scenes, generating many false positives in high-score region (high MR) and with a long processing time. For FCOS, AP performance of FeatureNMS is much higher than NMS, while JI and MR are similar. However for Faster RCNN, although FeatureNMS obtains a high recall, the others are worse than NMS, indicating that the feature embeddings trained with the cropped features in two-stage detectors are not distinctive enough, and there are many false positives in detections. Note that the learned embeddings in FeatureNMS have no explicit meaning except the relative distance between each pair, while ODAM-NMS directly uses heat maps that offer explanations of the detector model. With the default IoU threshold, our ODAM-NMS achieves better JI and MR than NMS and FeatureNMS for both detectors. Meanwhile, ODAM-NMS also achieves the best AP with Faster RCNN. The limitation of ODAM-NMS is that generating heat maps for dense predictions takes slightly longer (see §6.3.4). Overall, these results verify the object discrimination interpretation ability of ODAM with ODAM-Train and demonstrate that the instance-level explanation for predictions can help improve NMS in crowd scenes.

6.3.3 Ablation Study

We next provide the ablation studies evaluating NMS and ODAM-NMS with and without using ODAM-Train. The results are shown in the Tab. 12. When using the classical NMS, the baseline detector model yields similar and comparable results with and without using ODAM-Train, which demonstrates that ODAM-Train will have little influence on the baseline model performance when improving the explanation ability. This is a desirable property since we hope that providing explanations will not hinder the detector performance. As for ODAM-NMS, ODAM-Train brings an obvious improvement with the same parameter setting ($T^h = 0.8$ and $T_l = 0.2$) as compared to without ODAM-Train. This shows that model can be trained to give more consistent and separated explanations that benefit duplicate detection removal.

Furthermore, without ODAM-Train, ODAM-NMS needs the stricter judgment conditions to make more reasonable decisions, such as increasing T^h and reducing T_l, to obtain better results. This ablation study indicates that, since ODAM-NMS makes decisions based on heat-map correlations, it relies on good quality explanations, and using ODAM-Train is beneficial because it encourages the model to produce consistent and distinctive heat maps on detections of the same or different object.

6.3.4 Efficient calculation and inference time

Since there are many predictions from the object detector, calculating gradients of each prediction w.r.t. the feature maps one-by-one will incur an unacceptably long time cost for ODAM-Train and ODAM-NMS. To enable efficient ODAM-Train and ODAM-NMS, we adopt the RoI-pool features in the two-stage detector as A_k, since the gradients w.r.t. this layer for all predictions can be computed in a batch with the "autograd" function in PyTorch. As for one-stage detectors, the output features from the Feature Pyramid Network (FPN) [69] are adopted, and their gradients are computed in a batch through expansion of the gradient calculations [91] into a "reversed" detector head. This substantially improves the efficiency of ODAM-Train and ODAM-NMS.

The inference times are shown in Table 10. The inference time of vanilla NMS is 0.114s per image (on an RTX 3090 GPU),

while the inference time of ODAM-NMS is 0.178s. Thus, the extra calculation of the ODAM-NMS takes about 0.064s per image. Concretely, in the detector forward stage, 0.0072s is used for generating ODAM maps for predictions, and the ODAM-based NMS needs 0.0536s, compared with 0.0206s for vanilla NMS.

7 Conclusion

In this paper, we propose ODAM, a white-box gradient-based instance-level visual explanation technique for interpreting the predictions of object detectors. ODAM can produce instance-specific heat maps for any prediction attribute, including object class and bounding box coordinates, to show the important regions that the model uses to make its prediction. Our method is general and applicable to one- and two-stage and transformer-based detectors with different detector backbones and heads. Qualitative and quantitative evaluations demonstrate the advantages of our method compared with the class-specific (Grad-CAM) or black-box works (D-RISE), in terms of both object specification and object discrimination. We also conduct user trust studies and examine the similarity between human attention and XAI heat maps, and show that XAI heat maps that are more trustworthy also have higher similarity to human attention. Such a relationship provides a path for improving user trust of models by making the detector better mimic the human explanations, which will be considered in our future work.

Leveraging the object specification and object discrimination ability of ODAM, we propose two downstream applications for ODAM explanations. First, exploiting the object specification ability, we propose ODAM-KD, which performs knowledge distillation by combining the top-down attention from ODAM with bottom-up attention. Second, exploiting the object discrimination ability, we propose ODAM-Train and ODAM-NMS, which trains the detector to use more discriminative strategies and uses the visual explanations to infer duplicate detections for removal. In experiments, both proposed methods improve on the baseline performance, demonstrating the efficacy of ODAM on downstream applications. Note that such applications using visual explanations are not feasible with D-RISE due to its inefficiency. The successful applications also verify the interpretation ability of ODAM on detectors. By analyzing detectors with ODAM, we will continue to explore the possible directions for boosting object detection.

Acknowledgements

We thank Alice Yang for help with collecting the eye gaze data. This work was supported by grants from Collaborative Research Fund $\#C7129 - 20G$, Research Grant Council Hong Kong.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[2] A. Agrawal, D. Batra, and D. Parikh, "Analyzing the behavior of visual question answering models," *arXiv preprint arXiv:1606.07356*, 2016.

[3] D. Hoiem, Y. Chodpathumwan, and Q. Dai, "Diagnosing error in object detectors," in *ECCV*. Springer, 2012, pp. 340–353.

[4] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*. PMLR, 2017, pp. 1885–1894.

[5] 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 *ICCV*, 2017, pp. 618–626.

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?" explaining the predictions of any classifier," in *SIGKDD*, 2016, pp. 1135–1144.

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

[9] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421*, 2018.

[10] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *ICCV*, 2017, pp. 3429–3437.

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

[12] 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 *CVPR Workshops*, 2020, pp. 24–25.

[13] H. Wang, R. Naidu, J. Michael, and S. S. Kundu, "Ss-cam: Smoothed score-cam for sharper visual feature localization," *arXiv preprint arXiv:2006.14255*, 2020.

[14] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *CVPR*, 2018, pp. 8779–8788.

[15] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *CVPR*, 2017, pp. 7206–7215.

[16] S. A. Bargal, A. Zunino, D. Kim, J. Zhang, V. Murino, and S. Sclaroff, "Excitation backprop for rnns," in *CVPR*, 2018, pp. 1440–1449.

[17] 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 *CVPR*, 2021, pp. 11 443–11 452.

[18] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *ICCV*, 2019, pp. 9627–9636.

[19] Z. Chenyang and A. B. Chan, "Odam: Gradient-based instance-specific visual explanations for object detection," in *ICLR*, 2023.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[21] R. Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You

only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[26] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021, pp. 568–578.

[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*. Springer, 2020, pp. 213–229.

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

[29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*. Springer, 2014, pp. 818–833.

[30] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local explanation methods for deep neural networks lack sensitivity to parameter values," *arXiv preprint arXiv:1810.03307*, 2018.

[31] B. Kim, J. Seo, S. Jeon, J. Koo, J. Choe, and T. Jeon, "Why are saliency maps noisy? cause of and solution to noisy saliency maps," in *ICCVW*, 2019.

[32] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *arXiv preprint arXiv:1706.03825*, 2017.

[33] 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, p. e0130140, 2015.

[34] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *PR*, vol. 65, pp. 211–222, 2017.

[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, 2018.

[36] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *CVPR*, 2021, pp. 782–791.

[37] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.

[38] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *NeurIPS*, vol. 30, 2017.

[39] C.-H. Chang, E. Creager, A. Goldenberg, and D. Duvenaud, "Explaining image classifiers by counterfactual generation," *arXiv preprint arXiv:1807.08024*, 2018.

[40] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and fine-grained visual explanations for convolutional neural networks," in *CVPR*, 2019, pp. 9097–9107.

[41] J. Lee, J. Yi, C. Shin, and S. Yoon, "BBAM: Bounding box attribution map for weakly supervised semantic and instance segmentation," in *CVPR*, 2021, pp. 2643–2652.

[42] B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable basis decomposition for visual explanation," in *ECCV*, 2018, pp. 119–134.

[43] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, "Layercam: Exploring hierarchical class activation maps for localization," *TIP*, vol. 30, pp. 5875–5888, 2021.

[44] S. Srinivas and F. Fleuret, "Full-gradient representation for neural network visualization," *NIPS*, vol. 32, 2019.

[45] H. G. Ramaswamy *et al.*, "Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization," in *WACV*, 2020, pp. 983–991.

[46] A. Saha, A. Subramanya, K. Patil, and H. Pirsiavash, "Role of spatial context in adversarial robustness for object detection," in *CVPR Workshops*, 2020, pp. 784–785.

[47] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.

[48] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.

[49] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[50] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *CVPR*, 2019, pp. 2604–2613.

[51] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," *IEEE Trans. on Image Process*, vol. 28, no. 4, pp. 2051–2062, 2018.

[52] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[53] G. Chen, W. Choi, X. Yu, T. Han, and M. Chandraker, "Learning efficient object detection models with knowledge distillation," *NeurIPS*, vol. 30, 2017.

[54] Q. Li, S. Jin, and J. Yan, "Mimicking very efficient network for object detection," in *CVPR*, 2017, pp. 6356–6364.

[55] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou, "General instance distillation for object detection," in *CVPR*, 2021, pp. 7842–7851.

[56] T. Wang, L. Yuan, X. Zhang, and J. Feng, "Distilling object detectors with fine-grained feature imitation," in *CVPR*, 2019, pp. 4933–4942.

[57] R. Sun, F. Tang, X. Zhang, H. Xiong, and Q. Tian, "Distilling object detectors with task adaptive regularization," *arXiv preprint arXiv:2006.13108*, 2020.

[58] J. Guo, K. Han, Y. Wang, H. Wu, X. Chen, C. Xu, and C. Xu, "Distilling object detectors via decoupled features," in *CVPR*, 2021, pp. 2154–2164.

[59] L. Zhang and K. Ma, "Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors," in *ICLR*, 2020.

[60] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan, "Focal and global knowledge distillation for detectors," in *CVPR*, 2022, pp. 4643–4652.

[61] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[62] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *ICCV workshops*, 2019, pp. 0–0.

[63] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-nms–improving object detection with one line of code," in *ICCV*, 2017, pp. 5561–5569.

[64] S. Liu, D. Huang, and Y. Wang, "Adaptive nms: Refining pedestrian detection in a crowd," in *CVPR*, 2019, pp. 6459–6468.
[65] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *CVPR*, 2018, pp. 3588–3597.
[66] N. O. Salscheider, "Featurenms: Non-maximum suppression by learning feature embeddings," in *ICPR*. IEEE, 2021, pp. 7848–7854.
[67] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[69] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *CVPR*, 2017, pp. 2117–2125.
[70] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.
[71] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd," *arXiv preprint arXiv:1805.00123*, 2018.
[72] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *ICCV*, 2021.
[73] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.
[74] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.
[75] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.
[76] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
[77] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw Learn Syst*, vol. 28, no. 11, pp. 2660–2673, 2016.
[78] J. A. Oramas Mogrovejo, K. Wang, and T. Tuytelaars, "Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks," in *ICLR*, 2019.
[79] R. Qi, Y. Zheng, Y. Yang, J. Zhang, and J. H. Hsiao, "Individual differences in explanation strategies for image classification and implications for explainable ai." in *CogSci*, 2023.
[80] R. Qi, Y. Zheng, Y. Yang, C. C. Cao, and J. H. Hsiao, "Explanation strategies for image classification in humans vs. current explainable ai," *arXiv preprint arXiv:2304.04448*, 2023.
[81] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," in *AAAI*, vol. 32, no. 1, 2018.
[82] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.
[83] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *CVPR*, 2017, pp. 4133–4141.
[84] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *AAAI*, 2019, pp. 3779–3787.
[85] Y. Liu, J. Cao, B. Li, C. Yuan, W. Hu, Y. Li, and Y. Duan, "Knowledge distillation via instance relationship graph," in *CVPR*, 2019, pp. 7096–7104.
[86] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.
[87] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *ICCV*, 2019, pp. 1365–1374.
[88] X. Li, J. Wu, H. Fang, Y. Liao, F. Wang, and C. Qian, "Local correlation consistency for knowledge distillation," in *ECCV*. Springer, 2020, pp. 18–33.
[89] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *ICCV*, 2021, pp. 5311–5320.
[90] X. Chu, A. Zheng, X. Zhang, and J. Sun, "Detection in crowded scenes: One proposal, multiple predictions," in *CVPR*, 2020, pp. 12 214–12 223.
[91] Z. Liu and A. Chan, "Boosting adversarial robustness from the perspective of effective margin regularization," in *BMVC*, 2022.

Chenyang Zhao received the B.Eng. degree in Electrical Engineering from Xiamen University, Xiamen, China, and M.S. degree in Computer Science from School of Electronic and Computer Engineering, Peking University, Shenzhen, China, in 2016 and 2019, respectively. She is currently working towards the Ph.D. degree in Computer Science at the City University of Hong Kong. Her research interests include explainable AI and object detection.

Janet H. Hsiao received her Ph.D. in Informatics from University of Edinburgh and was a postdoctoral researcher at University of California San Diego. She is currently Head and Associate Professor in Department of Psychology, a Principal Investigator of the State Key Laboratory of Brain and Cognitive Sciences, and a Steering Committee member of Institute of Data Science at University of Hong Kong. She is also a Fellow of the Cognitive Science Society and serves on the Governing Board. Her research interests are in learning and visual cognition, including face recognition, reading, and object detection and identification.

Antoni B. Chan received the B.S. and M.Eng. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000 and 2001, and the Ph.D. degree in electrical and computer engineering from the University of California, San Diego (UCSD), San Diego, in 2008. He is currently a Professor in the Department of Computer Science, City University of Hong Kong. His research interests include computer vision, machine learning, pattern recognition, and music analysis.