DiffCAM: Data-Driven Saliency Maps by Capturing Feature Differences

Supplementary Material

A. Derivation for Solving w

The goal is to solve the optimal $\mathbf{w}$ that maximizes $J(\mathbf{w})$. We take the gradient of $J(\mathbf{w})$ w.r.t. $\mathbf{w}$ and set it to zero as

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0, \tag{8}$$

which leads to

$$\begin{aligned}
&(\mathbf{w}^T\mathbf{S_v}\mathbf{w})\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{S}_\mu\mathbf{w}) - (\mathbf{w}^T\mathbf{S}_\mu\mathbf{w})\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{S_v}\mathbf{w}) = 0 \\
&\Rightarrow (\mathbf{w}^T\mathbf{S_v}\mathbf{w})2\mathbf{S}_\mu\mathbf{w} - (\mathbf{w}^T\mathbf{S}_\mu\mathbf{w})2\mathbf{S_v}\mathbf{w} = 0 \\
&\Rightarrow \frac{\mathbf{w}^T\mathbf{S_v}\mathbf{w}}{\mathbf{w}^T\mathbf{S_v}\mathbf{w}}\mathbf{S}_\mu\mathbf{w} - \frac{\mathbf{w}^T\mathbf{S}_\mu\mathbf{w}}{\mathbf{w}^T\mathbf{S_v}\mathbf{w}}\mathbf{S_v}\mathbf{w} = 0 \\
&\Rightarrow \mathbf{S}_\mu\mathbf{w} - J(\mathbf{w})\mathbf{S_v}\mathbf{w} = 0 \\
&\Rightarrow {\mathbf{S_v}}^{-1}\mathbf{S}_\mu\mathbf{w} - J(\mathbf{w})\mathbf{w} = 0.
\end{aligned} \tag{9}$$

Considering $J(\mathbf{w})$ is a scalar, we denote it as λ, the above equation is equivalent to

$$\begin{aligned}
\lambda\mathbf{w} &= {\mathbf{S_v}}^{-1}\mathbf{S}_\mu\mathbf{w} \\
&= {\mathbf{S_v}}^{-1}(\mathbf{z}-\hat{\mu})(\mathbf{z}-\hat{\mu})^T\mathbf{w} \\
&= {\mathbf{S_v}}^{-1}(\mathbf{z}-\hat{\mu})\alpha,
\end{aligned} \tag{10}$$

where α is another scalar. Therefore, the optimal solution $\mathbf{w}^*$ is along the direction of ${\mathbf{S_v}}^{-1}(\mathbf{z}-\hat{\mu})$.

B. More Results on MNIST

In Figure 11, we present more randomly selected results on MNIST to demonstrate the effects of discriminative visual attention. Column 2nd-5th show saliency maps generated by DiffCAM and the baseline methods answering the "Why is" question. The last three columns present counterfactual saliency maps to explain "Why not".

C. More Results on ImageNet

In Figure 12, we present more randomly selected results on ImageNet. DiffCAM exhibits generally more accurate objection localization than the baselines. While most methods perform similar on some easier examples (e.g., large animals), DiffCAM shows advantages in excluding noisy backgrounds from the saliency map especially on harder cases. For example, for the television image in the 4th row, DiffCAM successfully recognizes the object region although the model predicts the right class with a low probability.

D. Supplementary Material for Counterfactual Explanation Experiments

D.1. Experimental Settings

(1) Dataset

CUB-200-2011 [54] is a bird dataset with 200 categories, annotated with part location data. It includes a total of 15 parts, namely Back, Beak, Belly, Breast, Crown, Forehead, Left Eye, Left Leg, Left Wing, Nape, Right Eye, Right Leg, Right Wing, Tail, and Throat. For each part, the ground truth per image is determined as the median pixel position labeled by five different Mechanical Turk users.

(2) Network

Given that ResNet-50, as a deep residual network architecture, has demonstrated exceptional performance and robustness across a wide range of computer vision tasks, and is widely used as a benchmark model in the XAI field, we chose ResNet-50 as the baseline model. Counterfactual explanations were generated using the output from its final convolutional layer.

(3) Evaluation

For evaluation metrics, we basically align with GALORE. To assess the effectiveness of Counterfactual Explanation on the CUB-200-2011 dataset, which includes part annotations, the ground truth is represented by $\mathcal{G} = \{(p_i, a_i, b_i)\}_{i=1}^{N}$, where parts are denoted as p, and a and b refer to the target and counterfactual classes, respectively. Define Precision (P) and Recall (R) as

$$P = \frac{|\{i \mid \mathbf{p}_i \in \mathbf{r}, a_i = a, b_i = b\}|}{|\{k \mid \mathbf{p}_k \in \mathbf{r}\}|}$$

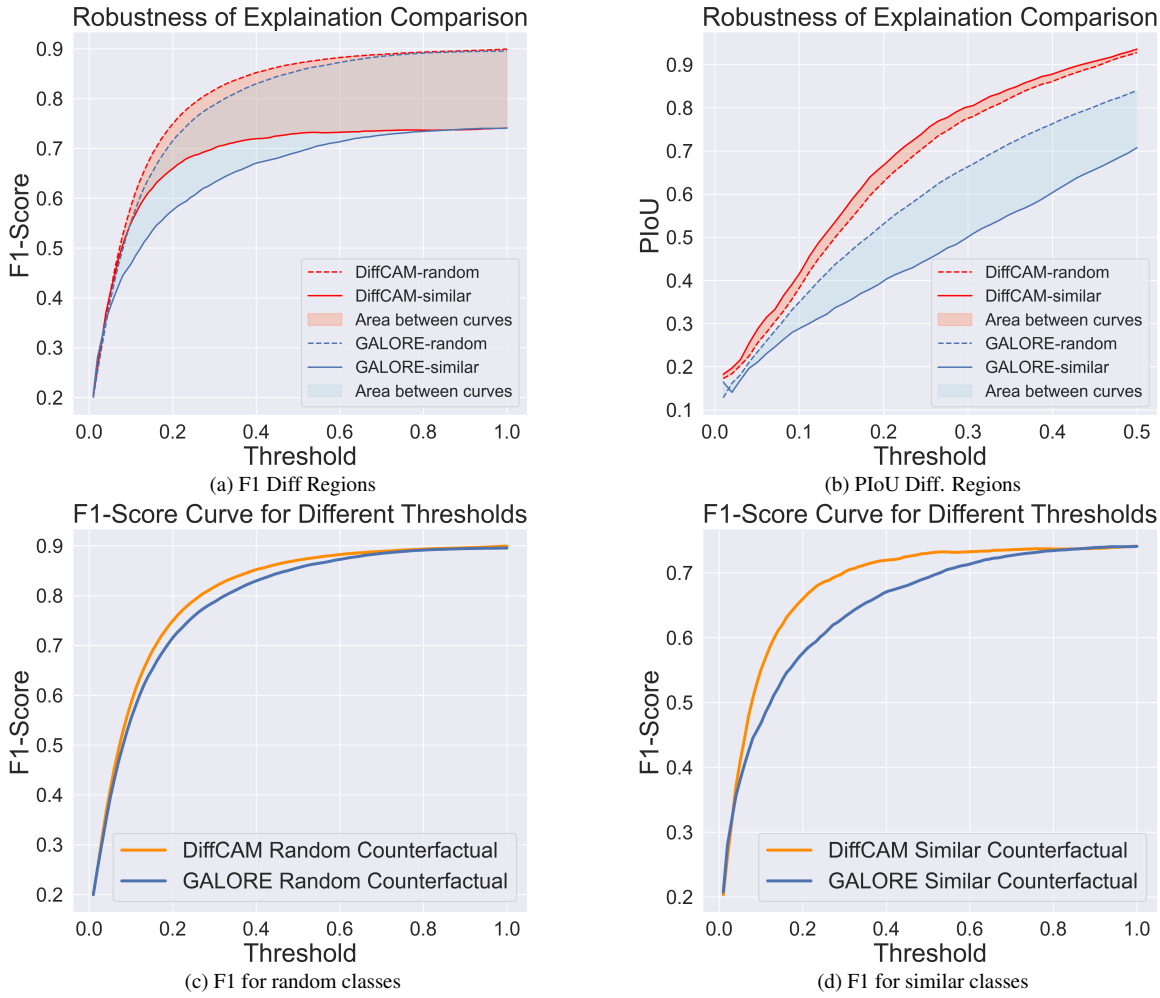
and

$$R = \frac{|\{i \mid \mathbf{p}_i \in \mathbf{r}, a_i = a, b_i = b\}|}{|\{i \mid (\mathbf{p}_i, a_i, b_i) \in \mathcal{G}, a_i = a, b_i = b\}|}$$

. Additionally, the F1-score

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}$$

is introduced to comprehensively compare model performance across different thresholds. Furthermore, Part Intersection over Union (PIoU) is defined to evaluate the semantic consistency of part-level regions by calculating the



(a) F1 Diff Regions

(b) PIoU Diff. Regions

(c) F1 for random classes

(d) F1 for similar classes

Figure 9. Evaluation of counterfactual explanations. (Comparison with GALORE)

overlap between part segments $r\{a,b\}(\mathbf{x})$ and $r\{b,a\}(\mathbf{x}^c)$, where k represents the number of part.

$$\text{PIoU} = \frac{|\{k \mid (p_k,a,b) \in r\{a,b\}(\mathbf{x})\} \cap \{k \mid (p_k,b,a) \in r\{b,a\}(\mathbf{x}^c)\}|}{|\{k \mid (p_k,a,b) \in r\{a,b\}(\mathbf{x})\} \cup \{k \mid (p_k,b,a) \in r\{b,a\}(\mathbf{x}^c)\}|}. \tag{11}$$

For a fair comparison, the heatmap size of the counterfactual regions in Table 2 is restricted to match the receptive field of a single unit, $\frac{1}{7*7} \approx 0.02$ of the area for ResNet-50.

D.2. More Results

Figure 10 displays the PIoU across different thresholds and counterfactual classes for the model. The results indicate that DiffCAM achieves better performance with similar classes, compared to random selected class. This is because DiffCAM maximizes the differences with similar classes, enabling it to generate heatmaps that focus more on features distinguishing the target class from similar classes. For classes with greater differences, the heatmaps emphasize the intrinsic features of the target class. Additionally, with a heatmap threshold set to 20% of the image size, DiffCAM reaches a PIoU value of 0.67.

E. Sanity Check Results

We use the ResNet-50 model pre-trained on ImageNet to perform sanity check. Specifically, we progressively randomize the parameters of the classification head (the fully connected layer) and the deep convolutional layers in the feature extractor. As shown in Figure 13, DiffCAM is unaffected by randomization in the classification head since its explanations are derived from deep features. However, it is highly sensitive to parameter perturbations in the feature extractor. These observations confirm that DiffCAM is closely tied to the model's behavior rather than generating explanations that are independent of the model and data.

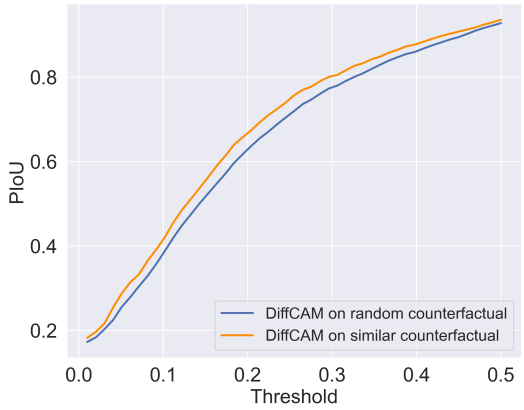


Figure 10. PIoU of DiffCAM under different counterfactual classes

F. Supplementary Material for Medical Imaging Experiments

F.1. Experimental Settings

F.1.1 Task Background

Image classification models are essential for many medical imaging applications, particularly in scenarios where fine-grained annotations are insufficient. Explanations such as saliency maps play a crucial role in building trust between AI systems and human experts. These saliency maps also provide insights of abnormality localization from deep models which automatically extract complex patterns in a data-driven way. In this study, we build upon previous work [7] by establishing an evaluation protocol using two chest X-ray datasets with fine-grained annotations. Note that despite the availability of segmentation or detection labels, we use them as ground truth for evaluating saliency maps. We simulate classification tasks using only image-level binary labels for model training.

F.1.2 Dataset Preparation

The RSNA Pneumonia Detection Challenge dataset [6] consists of chest X-ray images annotated with bounding boxes (bbox) to indicate potential areas of pneumonia. Each image is labeled with 1 (pneumonia present) or 0 (no pneumonia), and bounding boxes are provided only for pneumonia cases. The dataset includes a total of 14,863 images, split into 81% for training, 9% for validation, and 10% for testing. The training set contains 12,039 images from 4,870 patients with pneumonia, the validation set has 1,338 images from 541 patients with pneumonia, and the test set consists of 1,486 images from 601 patients with pneumonia.
The SIIM-ACR Pneumothorax Segmentation dataset [58] contains chest X-ray images with run-length-encoded (RLE) masks. Images with pneumothorax have associated masks, while those without do not. The dataset includes a total of 10,675 images, split similarly to the pneumonia dataset. The training set consists of 8,646 images from 1,931 patients with pneumothorax, the validation set has 961 images from 202 patients with pneumothorax, and the test set comprises 1,068 images from 246 patients with pneumothorax.

F.1.3 Model Training

For our study, we used an ImageNet-pretrained InceptionV3 model [11, 47] from Torch Hub as the base model. To adapt it for binary classification, we replaced the final fully connected layer with a two-class output layer. The model was fine-tuned on medical datasets with cross-entropy loss to improve diagnostic accuracy. Following the practice suggested in the previous work [7], training was conducted end-to-end with stochastic gradient descent (SGD), using a batch size of 64, an initial learning rate of 0.0001, momentum of 0.9, and weight decay of 0.00001. We selected the model checkpoint with the lowest validation loss to ensure optimal performance.

F.2. More Results

We present more examples of abnormality localization on the two medical imaging datasets. From Figure 14, DiffCAM delivers clearly more accurate localization on lung abnormalities for the RSNA dataset. For the SIIM-ACR dataset which is more challenging, we observe less alignment between the complex ground truth masks and the saliency maps generated by state-of-the-art XAI approaches. Nevertheless, we still notice the potential of DiffCAM in discovering discriminative differences between normal and abnormal X-ray images. For example, in the last row of SIIM-ACR images, only DiffCAM highlights the middle left region as part of key features relevant to pneumothorax.

G. More Visualization Results

G.1. DiffCAM on Confusing Classes.

Fig. 16 presents additional qualitative examples from CUB-200-2011 and ImageNet showing DiffCAM's capacity in providing accurate explanations among confusing classes. We note that (1) DiffCAM successfully highlights the glaucous-winged gull (col. 3), whereas GradCAM fails, and (2) DiffCAM correctly identifies the front legs and neck region as the key difference between the Eskimo Dog and the Timber Wolf (col. 4).

G.2. How DiffCAM Results Are Affected by Model and Reference Selection

We conducted the following experiments analyzing how DiffCAM results can be affected by difference choices of models and reference selections. Visualization results are shown in Fig. 17.

Models from strong to weak performance. We evaluate DiffCAM on three ImageNet pre-trained models, which are MobileNetV3, ResNet50 and ConvNeXt, with the top-1 accuracy of 68%, 76% and 84%, respectively. In line 1 of Fig. 17, we find stronger models generally yield better explanations on the multi-object image. The compared GradCAM result is generated with the best model ConvNeXt. DiffCAM is shown to outperform GradCAM with the same ConvNeXt model.

Impact of biased sample selection. In line 2 of Fig. 17, DiffCAM shows correct attention with different biased categories as references when answering the "why not" question. For example, when asking "why not water buffalo", DiffCAM highlights the airedale terrier which is similar to the result for "why is". However, when answering "why not Lakeland terrier", DiffCAM captures the cow as their most significant visual difference.

Impact of numbers of reference images. To capture reliable feature differences among classes, sufficient examples are typically required as the reference group. Nevertheless, DiffCAM is still flexible for a small number of or even one single reference example. As shown in line 3 of Fig. 17, given a car image, DiffCAM with only one nearest reference example captures the feature difference (the bridge) between two individual images, whereas using more references highlights the car body as common intra-class features.

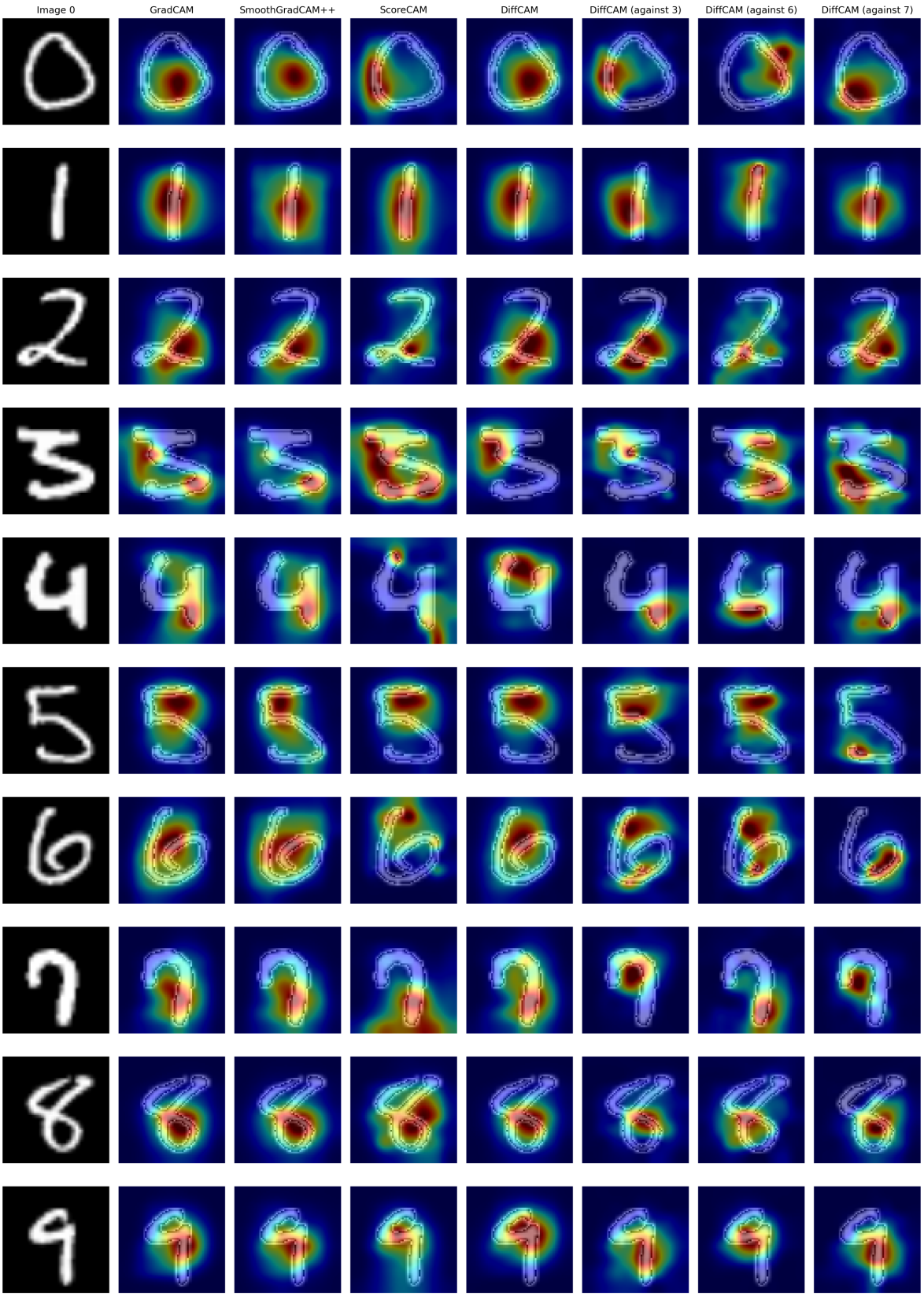


Figure 11. More examples of discriminative visual attention from the MNIST benchmark.

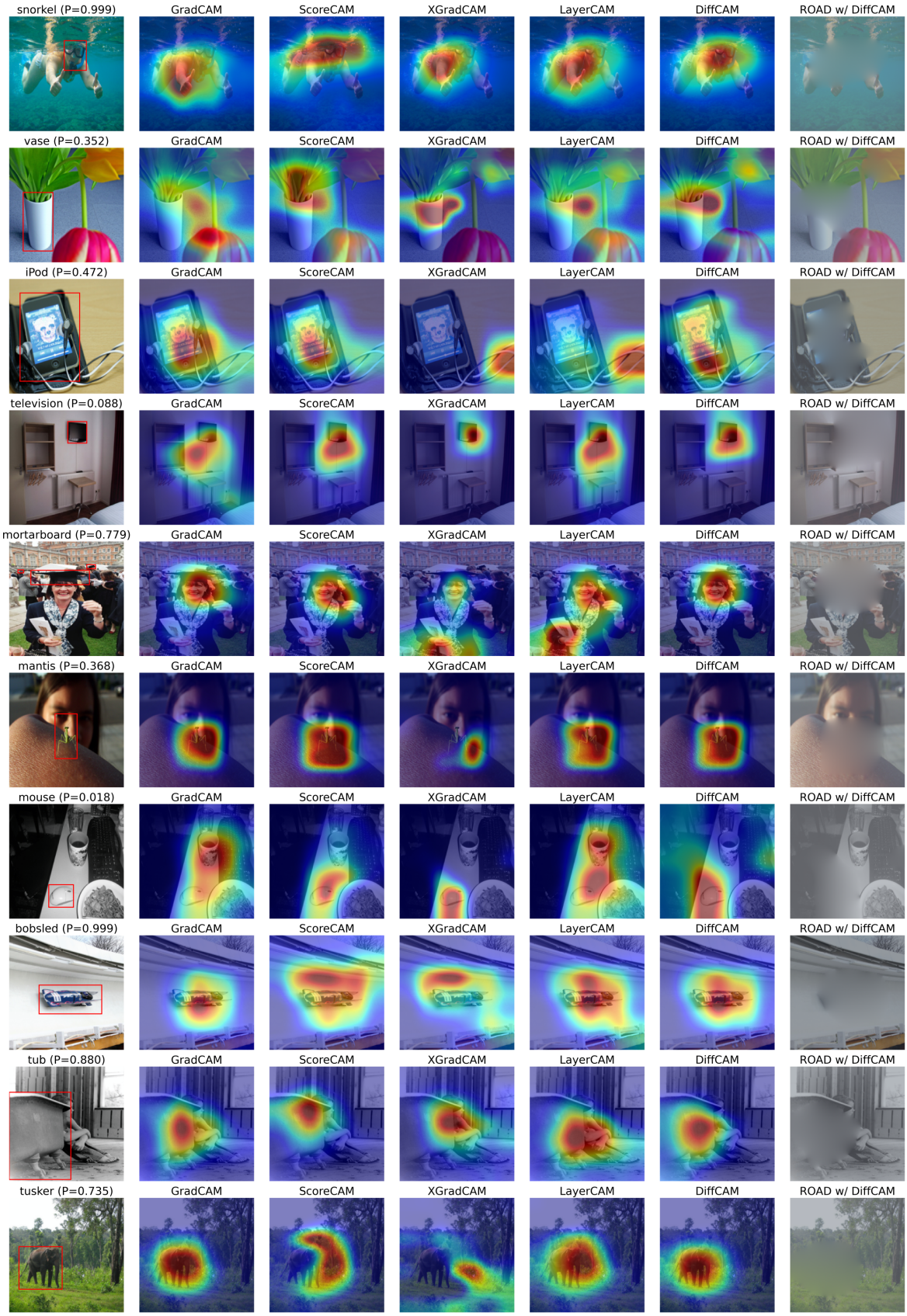


Figure 12. More examples of localization evaluation from the ImageNet benchmark.

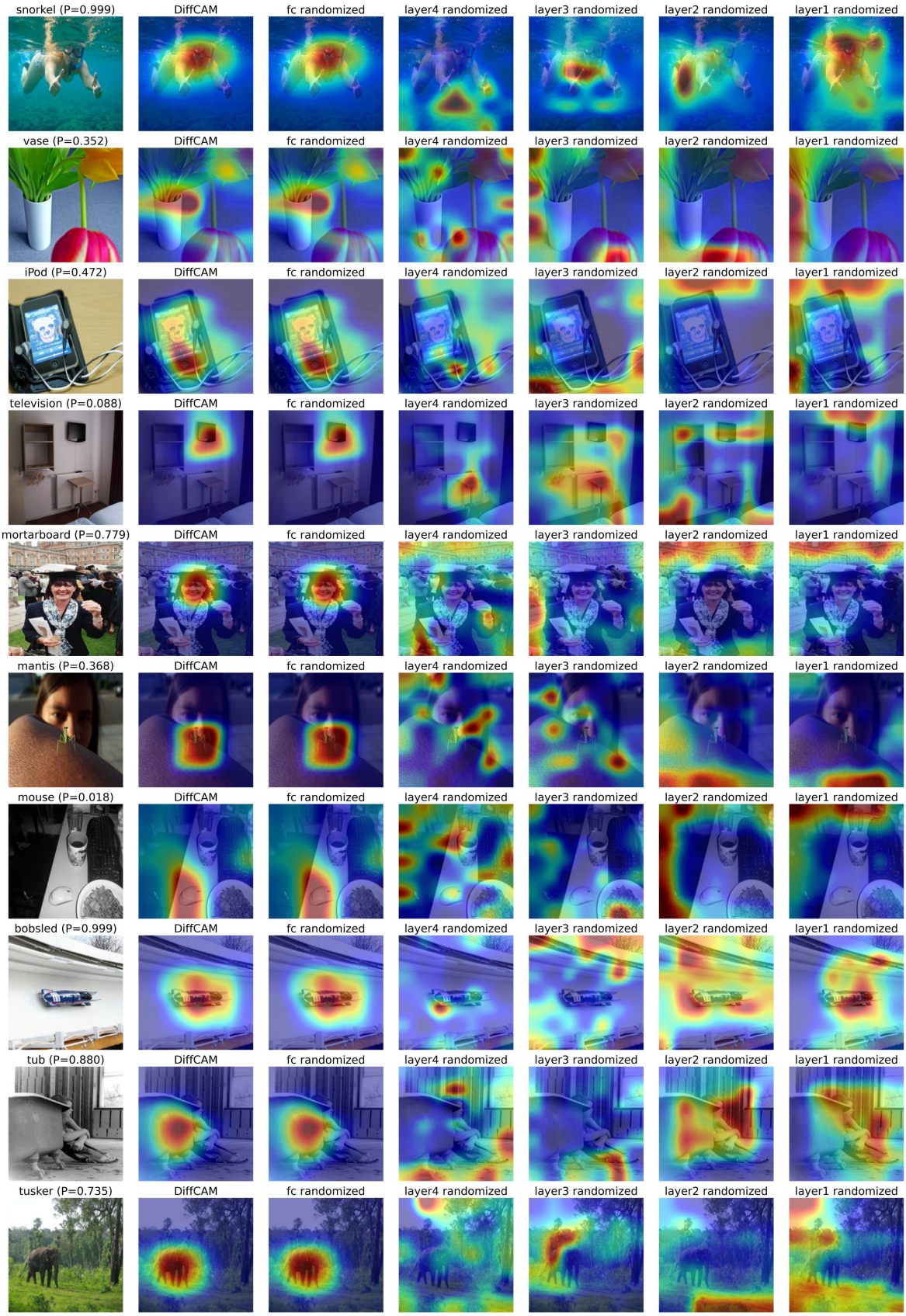


Figure 13. Sanity check on the ResNet-50 model pre-trained on ImageNet.

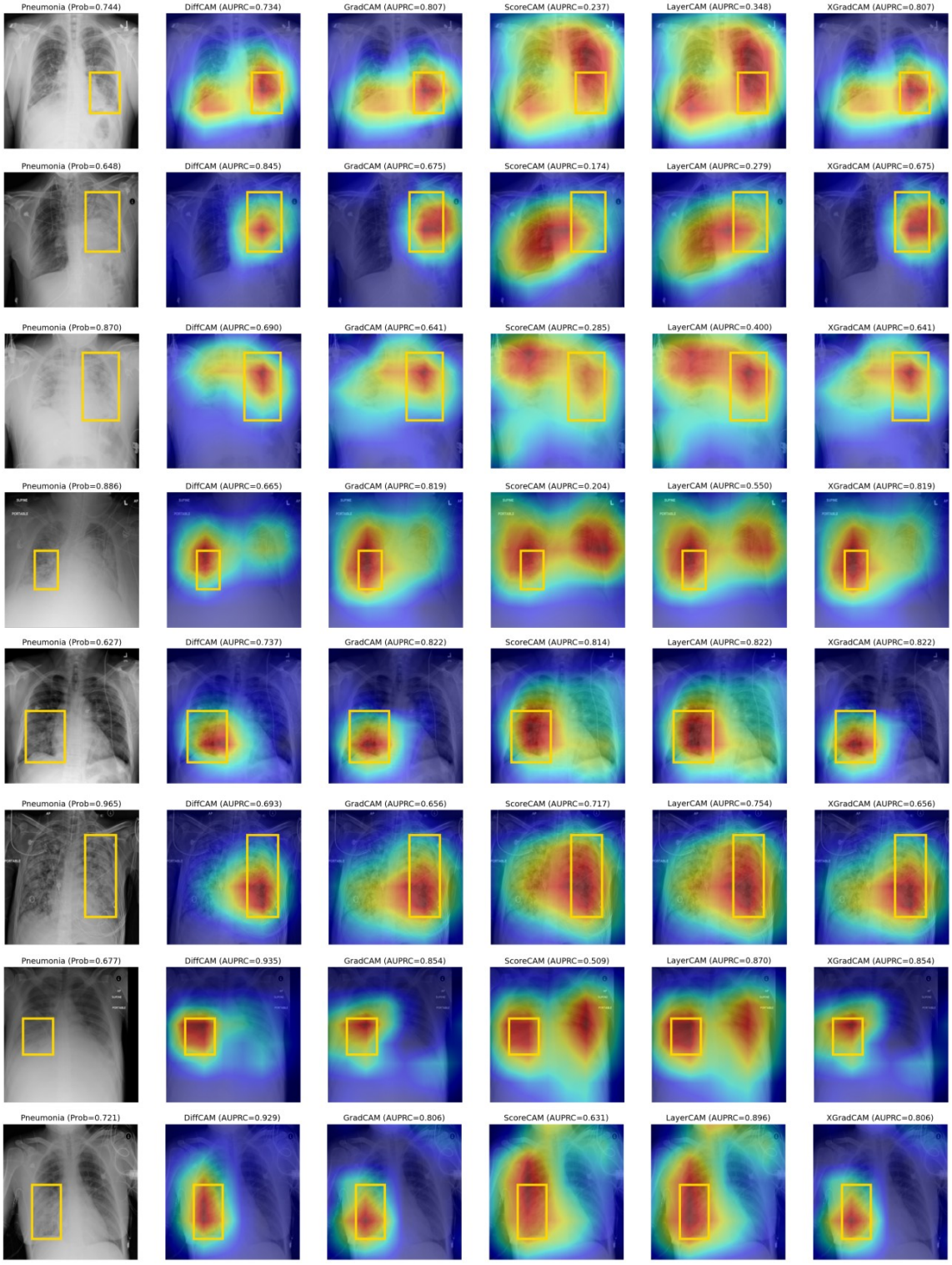


Figure 14. More examples of abnormality localization on the RSNA dataset.

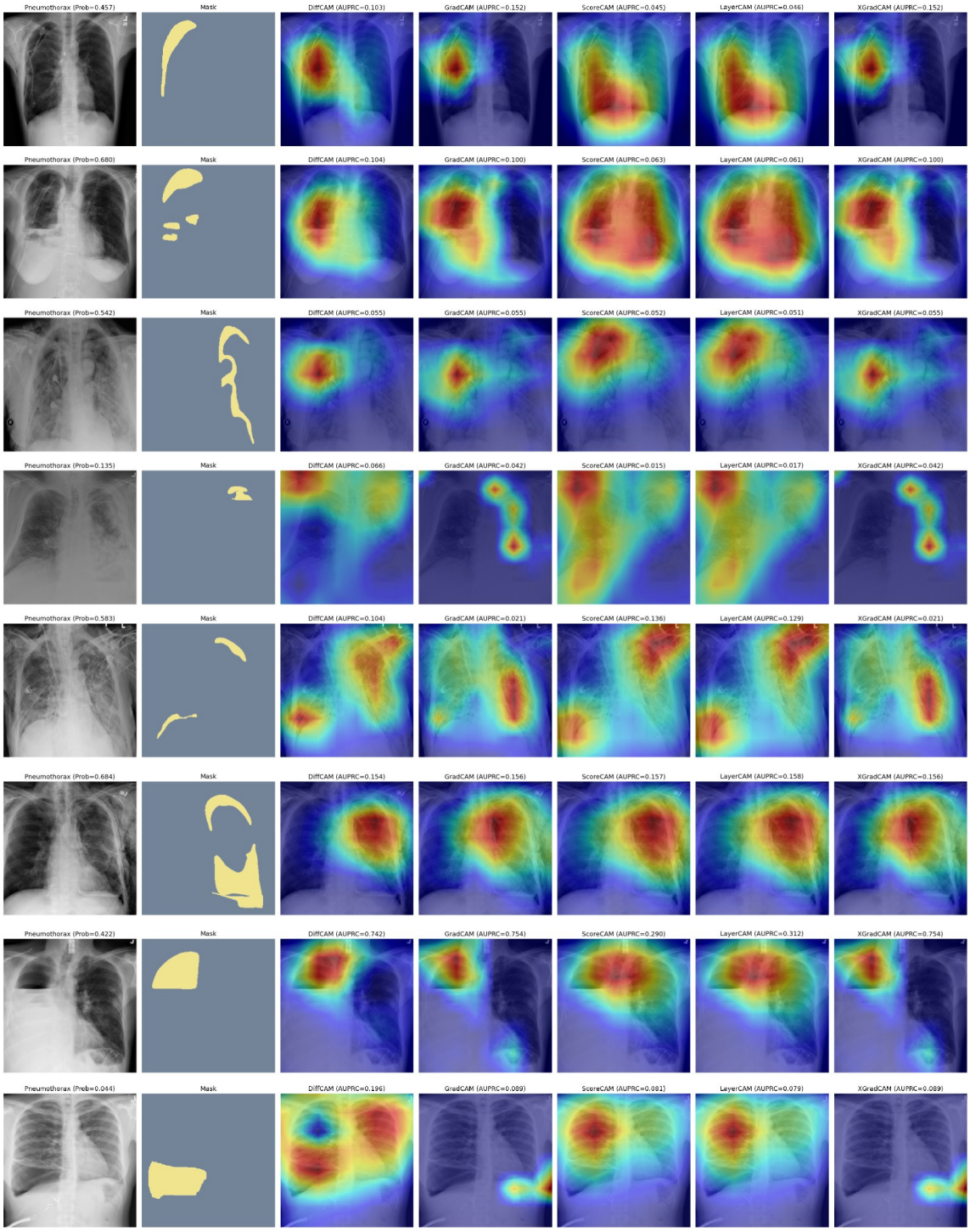


Figure 15. More examples of abnormality localization on the SIIM-ACR dataset.

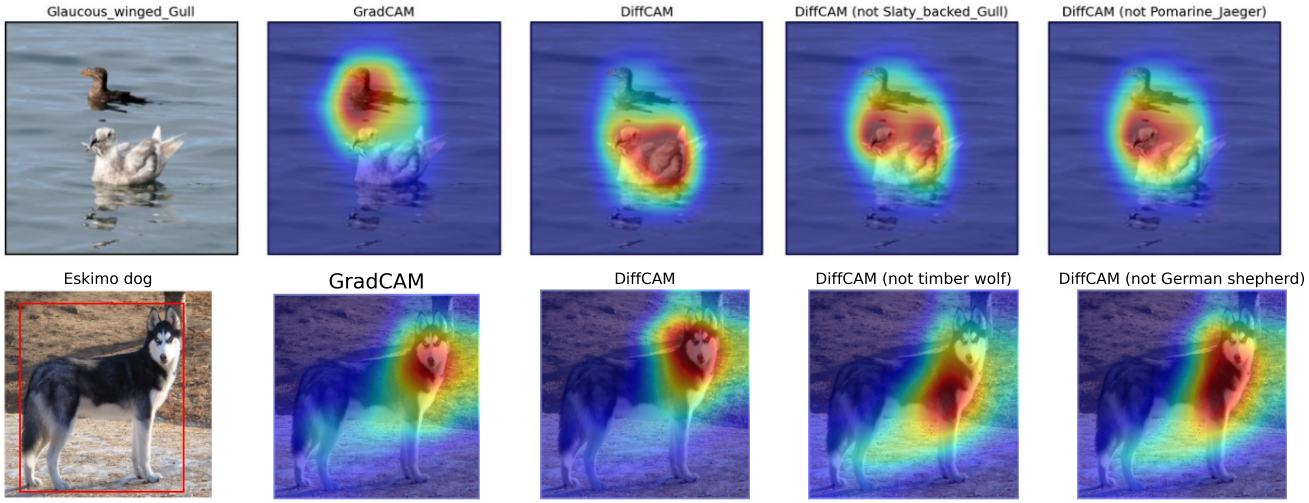


Figure 16. Qualitative examples on CUB-200-2011 and ImageNet.

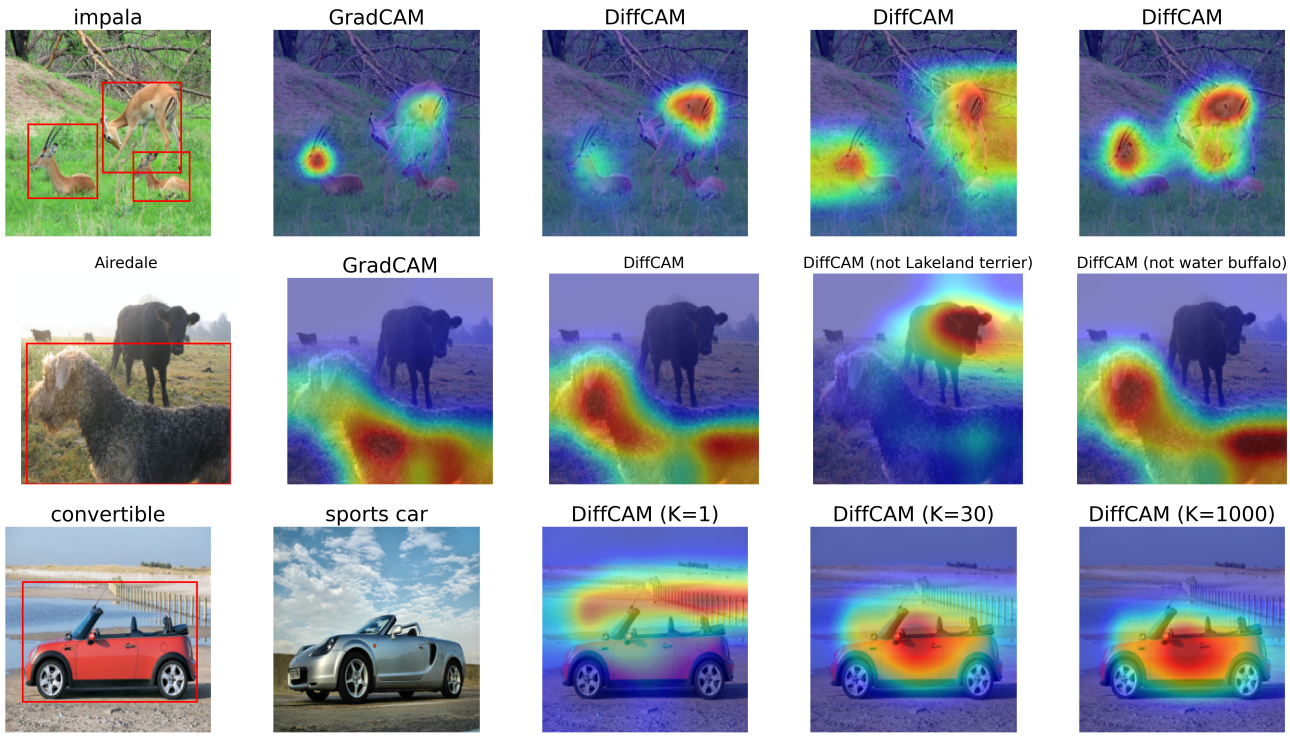


Figure 17. (1) DiffCAM on MobileNetV3, ResNet50 and ConvNeXt (Acc@1: 68%, 76%, 84%); (2) category-biased DiffCAM; (3) DiffCAM with different numbers of kNN reference samples.