

A-FMI: LEARNING ATTRIBUTIONS FROM DEEP NETWORKS VIA FEATURE MAP IMPORTANCE

Anonymous authors

Paper under double-blind review

ABSTRACT

Gradient-based attribution methods can aid in the understanding of convolutional neural networks (CNNs). However, the redundancy of attribution features and the gradient saturation problem, which weaken the ability to identify significant features and cause an explanation focus shift, are challenges that attribution methods still face. In this work, we propose: 1) an essential characteristic, Strong Relevance, when selecting attribution features; 2) a new concept, feature map importance (FMI), to refine the contribution of each feature map, which is faithful to the CNN model; and 3) a novel attribution method via FMI, termed **A-FMI**, to address the gradient saturation problem, which couples the target image with a reference image, and assigns the FMI to the “difference-from-reference” at the granularity of feature map. Through visual inspections and qualitative evaluations on the ImageNet dataset, we show the compelling advantages of A-FMI on its faithfulness, insensitivity to the choice of reference, class discriminability, and superior explanation performance compared with popular attribution methods across varying CNN architectures.

1 INTRODUCTION

As the understanding of neural networks is of crucial importance to engender user trust, interpreting network behavior has attracted increasing attention. To this end, attribution methods (Ancona et al., 2019) have demonstrated the remarkable ability in attributing the prediction of a given network, typically CNN, to its input. Regardless of various designs, a common axiom called completeness (Sundararajan et al., 2017) or local-faithfulness (Selvaraju et al., 2017) for most existing attribution methods can be summarized as: $\sum \text{attribution feature} \times \text{attribution score} \approx \text{network prediction}$. Simply stated, given an input image, attribution method determines the attribution score (also well-known as relevance (Bach et al., 2015) or contribution (Ribeiro et al., 2016)) of each attribution feature (*e.g.*, pixel, segment), in order to approximate the CNN’s prediction of interest. By redistributing such attributions to the input image, we can produce a saliency map to highlight the most important regions for predicting the class of interest.

In scrutinizing the axiom, we argue that a key challenge lies in the selection of attribution features, which should satisfy one essential characteristic — Strong Relevance (a statistic concept in feature selection (Blum & Langley, 1997)). Formally, a strong relevant feature is defined as a feature which has information that is both pertinent to the class prediction and can barely be derived from other features; in contrast, the redundant feature is highly correlated and can be represented by other features. Although the redundancy in attribution features does not affect the approximation of the prediction, it negatively influences the learning of the attribution scores. The key reason is that the entanglement among these features weakens the ability to identify significant features, and small changes of redundant features can swing the value of attribution scores widely.

Having realized the vital role of attribution feature selection, we categorize the prior works into three levels of granularity: **pixels** (Simonyan et al., 2014; Zeiler & Fergus, 2014; Springenberg et al., 2015; Bach et al., 2015; Shrikumar et al., 2016; Smilkov et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017; Lundberg & Lee, 2017), **regions** (Ribeiro et al., 2016; Zintgraf et al., 2017; Fong & Vedaldi, 2017; Dabkowski & Gal, 2017; Petsiuk et al., 2018; Fong et al., 2019; Kapishnikov et al., 2019), and **feature maps** (Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhyay et al., 2018). However, each type of methods suffers from some limitations:

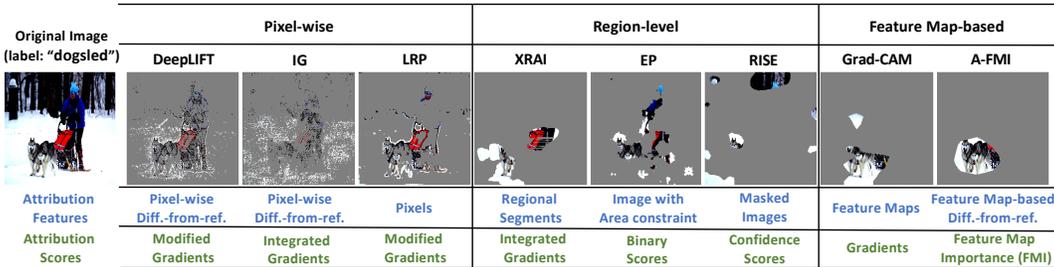


Figure 1: Visual comparison of present attribution methods and the proposed A-FMI at 10% of important pixels insertion for the prediction of class “dogsled”. Pixels with significant attributions are remained, while the rest is removed. A black reference is used for DeepLIFT, IG and A-FMI.

- One research line exploits single pixels as attribution features and learns the gradients, modified gradients or integrated gradients of the target class as the attribution scores. Despite great success, one limitation is that it violates the Strong Relevance characteristic of attribution features, since pixels are strongly related to their surrounding pixels. Such redundancy easily results in suboptimal and fragile attributions, being akin to the edge detector (Adebayo et al., 2018) and vulnerable to small perturbations (Ghorbani et al., 2019).
- Region-level methods measure how sensitive the prediction is to the masked image or perturbations of regional segments. However, they hardly satisfy the completeness principle, *i.e.*, adding the attributions of all regions up may not guarantee a match with the prediction. Moreover, built upon the perturbation or masking mechanism, they are time-consuming and heavily depend on the quality of segmentation.
- Decomposing the target prediction to the feature maps in the last convolutional layer via gradient learning has been studied. Compared with pixel-wise attribution methods that easily result in focus shift, feature map-based methods highlight the compact regions of interest (see Figure 1). We attribute such success to the Strong Relevance characteristic of feature maps, wherein only subtle correlations exist as shown in Yosinski et al. (2015); Shang et al. (2016). However, previous methods inherit gradient saturation problem, which underestimates the contributions and further leading untargeted explanations (see Figure 2).

In this work, feature map reference is introduced as a solution to the gradient saturation problems. More specifically, we implement a baseline input (*e.g.*, a black image), run a forward pass to get feature map references, and measure the differences with the original feature maps. To refine the contributions for feature map-based difference-from-references, we devise **feature map importance (FMI)**, which is a set of modified gradients estimated by the Taylor series. FMI allows significant feature maps to propagate signals even in situations where their gradients are zero. Moreover, FMI has notable properties (evidence in Section 4.1): 1) strong representation ability, embedding class-specific information which could be directly used to classify, 2) local faithfulness, *i.e.*, in the vicinity of the input image, FMI is locally accurate to the CNN.

We further propose a new **Attribution Method via FMI (A-FMI)**, which is equipped with difference-from-reference and FMI to interpret the CNN models. To validate the effectiveness of A-FMI, we perform thorough visual inspections and qualitative evaluations for VGG19 and ResNet50 on the ImageNet dataset. The empirical results demonstrate that A-FMI consistently produces better results than popular attribution methods and is conceptually advantageous in that: 1) compared with pixel-wise methods, it uses attribution features (*i.e.*, feature maps) with Strong Relevance characteristic, thus able to output compact regions of interest; moreover, distinct from other attribution methods who are sensitive to the selection of reference, A-FMI is more robust (evidence in Section 4.2); 2) compared with region-level methods, it satisfies the local faithfulness and is much more efficient; and 3) compared with feature map-based methods, it inherits the advantage of class discriminability (evidence in Section 4.3) and solves the gradient saturation problem, producing more refined and targeted explanations. In summary, this work makes the following contributions:

- We emphasize the important role of the selection of attribution features, and compare attribution features into three levels of granularity: pixels, regions, and feature maps *w.r.t.* the desirable Strong Relevance characteristic.

- To the best of our knowledge, we are the first to introduce the feature map reference and FMI into the attribution methods, to solve the gradient saturation problem.
- We conduct extensive experiments, demonstrating the effectiveness and consistency of A-FMI in interpreting VGG-based and ResNet-based architectures on the ImageNet dataset.

2 RELATED WORK

Pixel-wise attribution methods largely leverage the backpropagation way to redistribute the prediction through the whole CNN model to single pixels. For example, Gradient (Simonyan et al., 2014) and Input*Gradient (Shrikumar et al., 2016) use gradient of the prediction *w.r.t.* each pixel as attribution scores; DeconvNet (Zeiler & Fergus, 2014) and Guided Backpropagation (Springenberg et al., 2015) employ well-designed operations on the gradients of nonlinear activation functions. However, the gradient saturation problem is inherent in the backpropagation way, which easily results in the vanishing gradients and underestimating importance of pixels. To solve this problem, DeepLIFT (Shrikumar et al., 2017) employs a baseline (reference) image to calculate the modified partial derivatives of the difference-from-reference as the importance of pixels; meanwhile, Integrated Gradients (IG) (Sundararajan et al., 2017) aggregates the gradients by gradually varying the input from the baseline to the original image. Despite great success, the significant pixels highlighted by pixel-wise attribution methods are easily spread out — that is, the focus of the explanation model shifts into irrelevant edges, objects, or even background (see Figure 1, pixel-wise attribution methods select top 10% important pixels including snow and person for label dogsled).

Region-level attribution methods combine single pixels into patches or regions, and mainly apply the perturbation mechanism to directly evaluate the marginal effect of each region by masking or replacing it. For example, LIME (Ribeiro et al., 2016) approximates the CNN function by a sparse linear model between the patches and prediction, which is learned on perturbations of patches. Prediction Difference Analysis (Zintgraf et al., 2017) replaces patches with a sample from other images and obtain the contribution of each pixel by averaging the importance of patches containing the pixel. RISE (Petsiuk et al., 2018) randomly occludes an image and weights the changes in the confidence score of CNN. Extremal Perturbations (EP) (Fong et al., 2019) optimizes a spatial perturbation mask with a fixed area that maximally affects the CNN’s output. XRAI (Kapishnikov et al., 2019) incorporates the idea of gradients with the region-level features, coalescing smaller regions into larger segments based on the maximum gain of IG per region. These methods suffer from high computational complexity and heavily rely on the quality of segmentation.

Feature map-based attribution methods use feature maps in the last convolutional layer to produce saliency maps. Wherein, feature maps capture high-level semantic patterns and spatial information. CAM (Zhou et al., 2016) replaces the fully-connected layers with a global average pooling layer, and produces a coarse-grained saliency map via a weighted sum of feature maps. Furthermore, Grad-CAM (Selvaraju et al., 2017) uses the average gradients *w.r.t.* feature maps as the attribution scores. Though using gradients is beneficial, Grad-CAM inherits the gradient saturation problem, which is unexplored in the existing feature map-based methods. In this work, the proposed method, A-FMI, falls within this family and provides a potential solution to the saturation problem.

3 METHODOLOGY

Problem Formulation. Let F be the original CNN model that classifies an input image \mathbf{x} into one of C classes. The local explanation model designs to interpret $F(\mathbf{x})$ consisting of two components: attribution features and attribution scores. We utilize feature maps \mathcal{A} in the last convolutional layer as the attribution features, obtained by feeding \mathbf{x} into the CNN. Our goal is to learn the local explanation model (or instance-wise saliency map) $S^c(\mathcal{A})$ for a class of interest c that ensures local faithfulness: $\sum_{i,j} S^c(\mathcal{A})_{i,j} \approx F^c(\mathbf{x})$. In other words, the sum of entries in the saliency map should approximate the prediction of interest.

Gradient Saturation Problem. Exploiting the gradients of attribution features *w.r.t.* the target class as attribution scores is a prevalent technique in the existing attribution methods. However, the problem of gradient saturation is inherent in the gradient learning process, which easily leads to suboptimal attribution scores. The key reason is that, during the backward pass, vanishing gradients

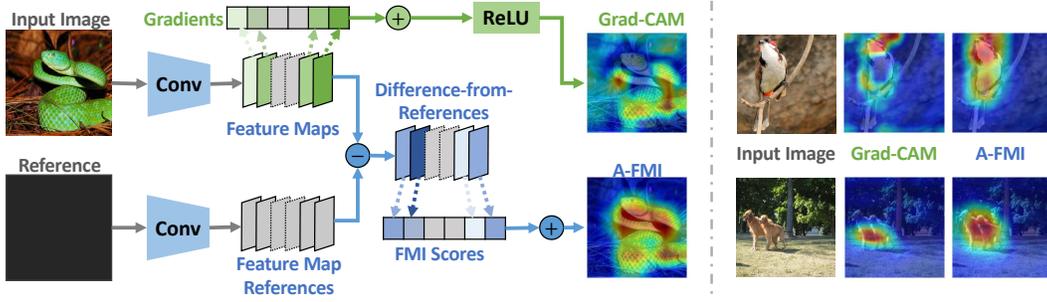


Figure 2: Overview (left) and Visual Comparisons (right) between Grad-CAM and A-FMI.

may occur due to nonlinear activation functions (*e.g.*, sigmoid, tanh, and even ReLU). Taking a two-layer ReLU network $F(\alpha) = \text{ReLU}(-\text{ReLU}(-\alpha + 1) + 2)$ as an example, the gradient of $F(\alpha)$ at $\alpha = 2$ is 0, intuitively indicating its trivial contribution. However, when changing α from 0 to 2, the network output changes from $F(\alpha = 0) = 1$ to $F(\alpha = 2) = 2$, which suggests that α is significant for prediction. Clearly, such vanishing gradients easily mislead the learning of attribution scores. Hence, It is of crucial importance to solve the gradient saturation problem.

Mathematical Motivation. For clarification of presentation, let us first consider a network that only consists of L fully connected layers as the classifier being interpreted, while first leaving the feature maps untouched. Let $\mathbf{x}^{(0)} \in \mathbb{R}^{d_0}$ denote the input vector associated with d_0 -dimensional features, and $\mathbf{y}^{(L)} \in \mathbb{R}^C$ be the prediction over C classes in the last layer (before softmax). In a forward pass, the output of each layer $\mathbf{x}^{(l)}$ is modeled as:

$$\mathbf{x}^{(l)} = \sigma(\mathbf{y}^{(l)}), \quad \mathbf{y}^{(l)} = \mathbf{W}^{(l)}\mathbf{x}^{(l-1)} + \mathbf{b}^{(l)} \quad (1)$$

where $\mathbf{W}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ and $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$ are the weight matrix and bias terms in the l -th layer; $\sigma(\cdot)$ is a nonlinear activation function; and d_l is the dimension of the l -th layer. Obviously, the j -th neuron in the l -th layer is represented using all neurons in the previous layer as: $x_j^{(l)} = \sigma(y_j^{(l)}) = \sigma(\sum_i W_{ji}^{(l)} x_i^{(l-1)} + b_j^{(l)})$. Hence, to measure the contribution of the i -th neuron in $(l-1)$ -th layer (*i.e.*, $x_i^{(l-1)}$) to the j -th neuron in the subsequent layer (*i.e.*, $x_j^{(l)}$), an intuitive solution is to leverage the partial derivative as: $\frac{\partial x_j^{(l)}}{\partial x_i^{(l-1)}} = \frac{\partial x_j^{(l)}}{\partial y_j^{(l)}} \cdot \frac{\partial y_j^{(l)}}{\partial x_i^{(l-1)}} = \sigma'(y_j^{(l)}) \cdot W_{ji}^{(l)}$. However, when $\sigma(y_j^{(l)})$ is flatten, $\sigma'(y_j^{(l)})$ approaches zero which may underestimate the contribution. To avoid unreasonable gradient vanishing, we employ the first order Taylor series to better estimate $\sigma'(y_j^{(l)})$ using a reference point $\bar{y}_j^{(l)}$, as:

$$\hat{\sigma}'(y_j^{(l)}) = \frac{\sigma(y_j^{(l)}) - \sigma(\bar{y}_j^{(l)})}{y_j^{(l)} - \bar{y}_j^{(l)}} \quad (2)$$

Here we use two cases to illuminate how the estimation process is of promise to solving the saturation problem: 1) When the activation function is ReLU, $\hat{\sigma}'(y_j^{(l)}) = \frac{\max\{0, y_j^{(l)}\} - \max\{0, \bar{y}_j^{(l)}\}}{y_j^{(l)} - \bar{y}_j^{(l)}}$. It always assigns a value in a range of $(0, 1)$ when the signs of $y_j^{(l)}$ and $\bar{y}_j^{(l)}$ are different; and it is zero if and only if $\sigma(y_j^{(l)}) = \sigma(\bar{y}_j^{(l)})$. 2) When the activation function is tanh, we set the reference as 0 for simplicity and have $\hat{\sigma}'(y_j^{(l)}) = \frac{1}{y_j^{(l)}} \int_0^{y_j^{(l)}} \sigma'(t) dt$, which is the average gradient of the nonlinearity in $[0, y_j^{(l)}]$. These estimators show the ability in identifying the significant features by assigning a non-zero value and filtering the irrelevant ones out by outputting a zero value.

Having established the estimator $\hat{\sigma}'(y_j^{(l)})$, we are able to approximate the difference of neuron $x_j^{(l)}$ from its own reference $\bar{x}_j^{(l)}$ in the l -th layer using all neurons in the $(l-1)$ -th layer as follows:

$$x_j^{(l)} - \bar{x}_j^{(l)} = \sum_i \hat{\sigma}'(y_j^{(l)}) \cdot w_{ji}^{(l)} \cdot (x_i^{(l-1)} - \bar{x}_i^{(l-1)}) \quad (3)$$

Consequently, we can define the modified gradient of the l -layer’s output $x_j^{(l)}$ w.r.t. $x_i^{(l-1)}$ as $\frac{\partial^\sigma x_j^{(l)}}{\partial x_i^{(l-1)}} = \sum_i \hat{\sigma}'(y_j^{(l)}) \cdot w_{ji}^{(l)}$. We further apply the chain rule of gradient to estimate the attribution of the original input $x_i^{(0)}$ to the final prediction $y_c^{(L)}$ of the last layer (before softmax) as:

$$\frac{\partial^\sigma y_c^{(L)}}{\partial x_i^{(0)}} = \sum_{p \in \mathcal{P}} (\prod \hat{\sigma}'_p \prod W_p) \quad (4)$$

where \mathcal{P} is the set of paths that connect the i -th neuron in the input (i.e., $x_i^{(0)}$) with the c -th neuron in the last layer (e.g., $y_c^{(L)}$). Iteratively, in the last layer, the difference between $y_c^{(L)}$ and its reference $\bar{y}_c^{(L)}$ can be successfully fetched by a weighted sum of all inputs:

$$y_c^{(L)} - \bar{y}_c^{(L)} = \sum_i \frac{\partial^\sigma y_c^{(L)}}{\partial x_i^{(0)}} \cdot (x_i^{(0)} - \bar{x}_i^{(0)}) \quad (5)$$

An image that has no specific information for any class, i.e., has a near-zero prediction score, can be a proper reference. This encourages the sum of attributions to approximate the prediction of interest. Moreover, unlike previous studies that mostly treat instance-wise baselines as hyper-parameters and need to carefully select them (e.g., adversarial examples or blurred images) to improve the explanation quality, we simply fix an identical reference for all images, to which our attribution scores are insensitive.

Attribution Method via FMI (A-FMI). Having obtained the potential solution to the gradient saturation problem, we subsume our feature map-based attribution method under this solution, as Figure 2 shows. Distinct from Grad-CAM that only takes the target image as input, we additionally couple it with a reference image (e.g., a black image) and feed them into the CNN model. The CNN model yields the feature maps $\mathcal{A} = \{\mathbf{A}^k\}$ with the feature map references $\bar{\mathcal{A}} = \{\bar{\mathbf{A}}^k\}$ in the last convolutional layer, and the outputs $\{\mathbf{x}^{(L)}\}$ and output references $\{\bar{\mathbf{x}}^{(L)}\}$ (before softmax) in the last fully-connected layers. Thereafter, to obtain the FMI_k^c of the k -th feature map $\mathbf{A}^k \in \mathcal{A}$ to the target prediction $y_c^{(L)}$, we simply change the single neuron $x_i^{(0)}$ in Equation equation 4 with the entry of \mathbf{A}_{ij}^k and average all modified gradients in the feature map \mathbf{A}^k , which is formulated as follows:

$$\text{FMI}_k^c = \frac{1}{N} \sum_i \sum_j \frac{\partial^\sigma y_c^{(L)}}{\partial A_{ij}^k} = \frac{1}{N} \sum_i \sum_j \sum_{p \in \mathcal{P}} (\prod \hat{\sigma}'_p \prod W_p), \quad (6)$$

where A_{ij}^k is the entry of \mathbf{A}^k located in the i -th row and j -th column; N is the total number of entries in the feature map \mathbf{A}^k ; and \mathcal{P} is the set of paths starting from any neuron in this feature map to the prediction score $y_c^{(L)}$. In essence, FMI_k^c is the modified weight of the fully connected layers in the CNN model, which directly captures the contribution of feature map \mathbf{A}^k in predicting target class c .

The saliency map $\mathbf{S}_{\text{A-FMI}}^c$ is modeled as a weighted linear combination of the difference between the feature maps $\{\mathbf{A}^k\}$ and feature map references $\{\bar{\mathbf{A}}^k\}$ with the corresponding FMI_k^c scores as the coefficients. More formally, the weighted combination can be represented as follows:

$$\mathbf{S}_{\text{A-FMI}}^c = \sum_k \text{FMI}_k^c \cdot (\mathbf{A}^k - \bar{\mathbf{A}}^k) \quad (7)$$

Note that feature map importance FMI_k^c differs from the importance weights in Grad-CAM.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following four research questions:

- **RQ1:** Are the explanations of A-FMI faithful to the CNN model?
- **RQ2:** How do different reference images affect the explanations of A-FMI?
- **RQ3:** Can A-FMI distinguish between different classes of interest?
- **RQ4:** How does A-FMI perform compared with other popular attribution methods?

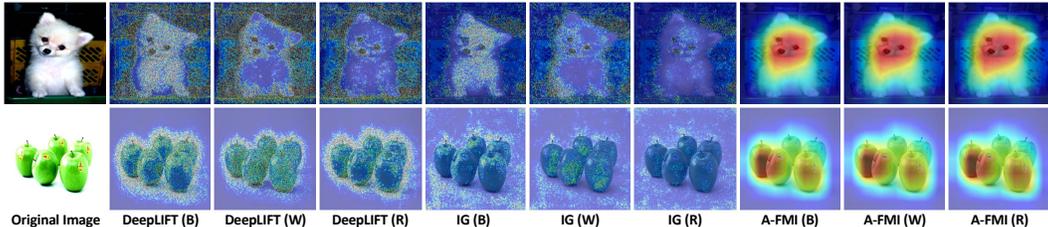


Figure 3: **Reference reliability.** A-FMI is reliable for the different choices of reference, while both DeepLIFT and IG determine the attribution strongly relative to a chosen reference. B, W, and R stand for black, white, and random reference respectively. See more samples in Appendix C.

To answer RQ1, we qualitatively evaluate the faithfulness of A-FMI to the VGG19 (Simonyan & Zisserman, 2015) model on the ImageNet (Russakovsky et al., 2015) validation set, as well as a simple CNN model (Paszke et al., 2019) on the MNIST (Yann et al., 1999) validation set, where only correctly predicted images are considered (*cf.* Section 4.1). To answer the remaining questions, we compare A-FMI to popular attribution methods: pixel-wise (Gradient, DeepLIFT, IG, LRP), region-level (XRAI, EP, RISE), and feature map-based (Grad-CAM). Without specification, we use a black image as the reference in DeepLIFT, IG, and A-FMI. Visual inspections on VGG19 and qualitative metrics on both VGG19 and ResNet50 (He et al., 2016) are used to validate the effectiveness of A-FMI in terms of reference reliability (*cf.* Section 4.2), class discriminability (*cf.* Section 4.3), and explanation quality (*cf.* Section 4.4). To ensure the reproducibility of our work, we have uploaded the code of A-FMI, all baselines, and their comparisons in the supplementary material.

4.1 FAITHFULNESS OF EXPLANATION (RQ1)

The faithfulness of an explanation model is its ability to accurately estimate the function learned by the CNN model. Typically, it can be described on two levels: 1) local faithfulness, justifying whether the explanations corresponds to the CNN predictions in the vicinity of an image instance, and 2) global faithfulness, validating whether the globally important features for a target class are identified. However, as the evaluation of global faithfulness is seldom performed in existing attribution methods, no measuring framework is available. Hence, we propose a measuring framework which uses FMI to classify the images to evaluate the faithfulness of A-FMI from both local and global perspectives.

To this end, for a specific image in class c , we use the attribution scores of its feature maps $FMI^c = \{FMI_1^c, \dots, FMI_K^c\}$ as the refined representation of the image. We then average the class-specific attribution scores over all training images as \overline{FMI}^c , which can be viewed as the representation of the prototype in class c . Thereafter, we classify each image in the validation set based on its cosine similarity with the prototype representations $\{\overline{FMI}^1, \dots, \overline{FMI}^C\}$. If the FMI-based class is identical to the target class c , the explanation of an image is faithful to the CNN model. Impressively, we achieve an average explanation accuracy of **86.9%** and **88.4%** in MNIST with 10 classes and ImageNet datasets with 1000 classes, respectively.

We present a faithfulness analysis of A-FMI based on the classification accuracy. First, the instance-wise FMI^c is well qualified to be a representation of an image, which encodes the information pertinent to the class c from a single image and hence reflects the local faithfulness. Then, the cosine similarity-based accuracy indicates that images with the same classes tend to form clusters, and the prototype representation \overline{FMI}^c captures the class-wise patterns. This further suggests that A-FMI to some extent achieves the global faithfulness.

4.2 REFERENCE RELIABILITY (RQ2)

To engender user trust, a reliable and trustworthy explanation model should be robust to factors that do not contribute to the model prediction. Hence, we explore how the reference, a factor additionally introduced to solve the saturation problem, affects DeepLIFT, IG, and A-FMI. Accordingly, we consider the variants of A-FMI, DeepLIFT, and IG that use different references — a black image, a white image, and an image filled with random pixels to produce saliency maps in Figure 8.

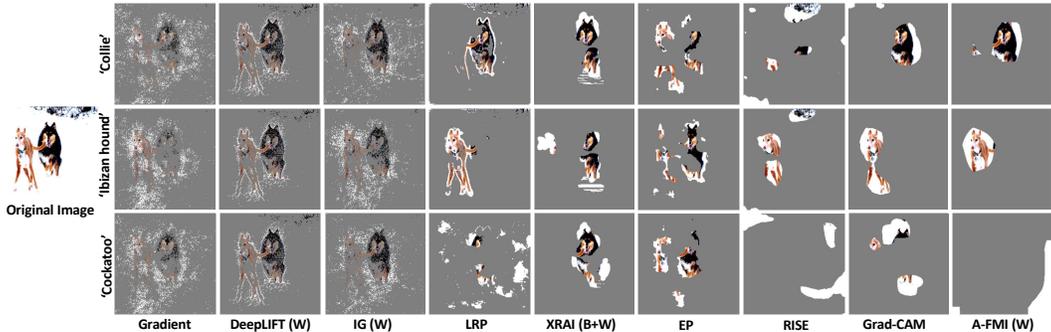


Figure 4: **Class discriminability.** Category-specific visualizations of all methods at a fixed 10% of important pixels insertion. The original image contains exactly two categories - Collie and Ibizan hound. The third category - Cockatoo - is the minimal activated category.

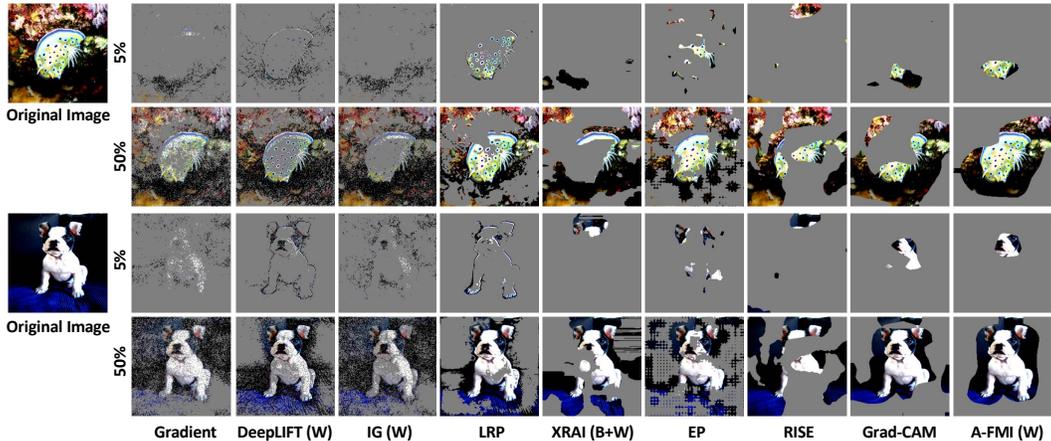


Figure 5: **Visual Inspections.** A visual comparison of all methods, where 5% and 50% important pixels are included, respectively. See more examples in Appendix D.

We observe that DeepLIFT and IG shift their focus when using different references. In particular, the explanations of DeepLIFT and IG with white reference tend to focus on the darker pixels and vice versa. For example, DeepLIFT(W) and IG(W) pay more attention to the black background than the white dog (1st row), while DeepLIFT(B) and IG(B) emphasize the white background surrounding the green apples (2nd row). Moreover, the results of DeepLIFT(R) and IG(R) are unsatisfactory when utilizing a random reference. This verifies that the selection of reference significantly affects the explanations of DeepLIFT and IG. In contrast, the saliency maps of A-FMI are more consistent, suggesting that the proper reference acts as a prior for A-FMI rather than a hyperparameter. This validates the robustness and reliability of A-FMI without hinging on instance-by-instance solutions (Kindermans et al., 2019).

4.3 CLASS DISCRIMINABILITY (RQ3)

A reasonable explanation method should be able to produce discriminative visualizations for different class of interest (Selvaraju et al., 2017). Figure 4 shows a category-specific visual comparison of all methods with 10% of insertion on an image with two classes: collie and Ibizan. We also display the visual explanation with a minimal activated class: cockatoo. Clearly, Gradient, DeepLIFT, IG, and XRAI hardly generate class-specific explanations, since the significant pixels only slightly change when different labels are assigned. A-FMI is able to output class discriminative explanations, which evidently shows that the relationships between input and predictions are successfully captured. Furthermore, we find that LRP is prone to outlining the edges even for background (*cf.* Figure 1), while Grad-CAM and A-FMI output tightly identify a region of interest in the image.

Table 1: Performance comparison *w.r.t.* area under curves of Accuracy.

Accuracy-AUC	Random	Gradient	DeepLIFT	IG	LRP	XRAI	EP	RISE	GradCAM	A-FMI*
VGG19	0.0854	0.1657	0.1673	0.2177	0.3460	0.3269	0.1775	0.2596	0.5343	0.5908*
ResNet50	0.0854	0.2236	0.1812	0.2857	-	-	0.2613	0.2970	0.6380	0.6513*
Time per image/s	-	0.0354	0.0669	1.5908	0.9719	14.6469	14.1818	10.8123	0.0291	0.1392

4.4 OVERALL PERFORMANCE COMPARISON (RQ4)

Visual Inspections. Figure 5 shows that: 1) For pixel-wise attribution methods, at 5% insertion, DeepLIFT and LRP initially capture outlines, while Gradient and IG fail to find any meaningful patterns. As the percentage of pixel insertion increases, Gradient, DeepLIFT, and IG become more defined, however, the salient pixels are largely distributed to unrelated areas; meanwhile, LRP tends to overvalue the lines, edges and corners of the images. 2) For region-level attribution methods, XRAI, EP and RISE output more compact regions. However, thier performances are unstable. 3) For feature map-based methods, Grad-CAM and A-FMI tend to initially search for the significant characteristics of target object first (eyes or nose for dog) and then expand outward (body for dog). Moreover, the explanations of A-FMI are more targeted than Grad-CAM, which suffers from the gradient saturation problem. This verifies the rationality and effectiveness of using references. Sufficient visualizations of all methods with thresholds that vary from 5% to 50% are provided in Appendix D.

Predictive Performance. *Insertion* metric is used to quantitatively evaluate the explanations. *Deletion* metric is not used here since Hooker et al. (2019) has shown its drawbacks. In particular, we start with a black image, gradually add the pixels with high confidence, and feed this masked image into the CNN model. As the percentage of pixel insertion increases from 0% to 100%, we monitor the changes to classification accuracy (*i.e.*, the fraction of masked images that are correctly classified). A sharp increase as well as a higher area under the accuracy curve indicate a better explanation. In this experiment, we focus on 5000 images sampled from ImageNet and show the results in Figure 9 and Table 1¹. We find that Grad-CAM and A-FMI consistently outperform the other attribution methods, being able to identify pixels that are truly important to CNN as those with the highest attributions chosen by the methods. We attribute this success to the Strong Relevance characteristic of feature maps. Moreover, in VGG19, A-FMI achieves significant improvements over Grad-CAM, indicating that using FMI with the reference is a promising solution to the saturation problem; whereas, A-FMI outperforms slightly better than Grad-CAM in ResNet50. This is reasonable since one fully connected layer is involved in ResNet50 and has a minor saturation problem. We also find that pixel-wise attribution methods perform relatively poor which is consistent to the visual inspections.

Computational Performance. In terms of time complexity, we report the time cost per image of each attribution method and find that, A-FMI performs similarly to the pixel-wise attribution methods while being significantly faster than the region-level attribution methods.

5 CONCLUSION

We proposed a novel attribution method via feature map importance, A-FMI, to produce visual explanations for CNNs. A-FMI provides a potential way to solve the gradient saturation problem at the granularity of feature maps, which allows the information to be backpropagated even when the gradient approaches zero. Extensive experiments illustrated the superior performance of A-FMI from both interpretable and faithful perspectives, compared with other popular attribution methods. Future work includes applying the attribution methods to other types of neural networks, such as graph neural networks.

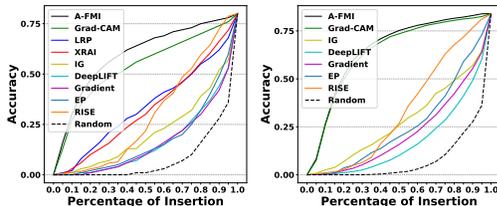


Figure 6: Accuracy curves of various attribution methods *w.r.t.* different percentage of important pixel insertion (Left: VGG19; Right: ResNet50).

¹The results are different from that reported in Petsiuk et al. (2018), where blurred canvases and probability curve are adopted for insertion.

REFERENCES

- Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, pp. 9525–9536, 2018.
- Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus H. Gross. Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700, pp. 169–191. 2019.
- Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.
- Avrim Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271, 1997.
- Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*, pp. 839–847, 2018.
- Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NeurIPS*, pp. 6967–6976, 2017.
- Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, pp. 2950–2958, 2019.
- Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pp. 3449–3457. IEEE Computer Society, 2017.
- Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *EAAI*, pp. 3681–3688, 2019.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.
- Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, pp. 9734–9745, 2019.
- Andrei Kapishnikov, Tolga Bolukbasi, Fernanda B. Viégas, and Michael Terry. XRAI: better attributions through regions. In *ICCV*, pp. 4947–4956, 2019.
- Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700, pp. 267–280. 2019.
- Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pp. 4765–4774, 2017.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.
- Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *BMVC*, pp. 151, 2018.
- Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144, 2016.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017.
- Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *ICML*, pp. 2217–2225, 2016.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, 2016. URL <http://arxiv.org/abs/1605.01713>.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, volume 70, pp. 3145–3153, 2017.
- Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR*, 2014.
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.
- Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *ICLR*, 2015.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, volume 70, pp. 3319–3328, 2017.
- Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pp. 5987–5995, 2017.
- LeCun Yann, Cortes Corinna, and Burges Christopher J.C. The mnist database of handwritten digits. 1999. URL <http://yann.lecun.com/exdb/mnist/>.
- Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas J. Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, 2015.
- Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, volume 8689, pp. 818–833, 2014.
- Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pp. 2921–2929, 2016.
- Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *ICLR*, 2017.

APPENDIX

In this appendix, we first report the reproducibility notes of A-FMI, including the implementation details, baselines, and evaluation metrics. In what follows, we present the visual inspections of A-FMI on different CNN Models to show the generalization ability. We then present more comparisons between A-FMI and attribution methods *w.r.t.* reference reliability and overall performance.

A REPRODUCIBILITY

A.1 IMPLEMENTATION DETAILS

Running Environment. The experiments are conducted on a single Linux server with 80 Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz and use a single NVIDIA Tesla V100-SXM2 32GB. Our proposed A-FMI is implemented in Python 3.7 and Pytorch 1.2.0, and we have provided the codes in the supplementary material.

Datasets. We use two datasets in Torchvision to validate the effectiveness of attribution methods: MNIST (Yann et al., 1999) and ImageNet (Russakovsky et al., 2015), in the context of image classification. We remain the training and validation sets as the original.

CNN models of Interest. On the MNIST dataset, we use the CNN example provided in the Pytorch tutorial² as the model being interpreted. On the ImageNet dataset, we use the pretrained CNNs in Torchvision³ as the models being interpreted, which include VGG-16 (Simonyan & Zisserman, 2015), VGG-19 (Simonyan & Zisserman, 2015), ResNet-50 (He et al., 2016), ResNet-152 (He et al., 2016), and ResNeXt-101-32x8d (Xie et al., 2017). These CNNs are fixed during the interpreting process.

A.2 BASELINES

We give detailed attribution methods as follows and include the implementations in the supplementary material. For each method, the hyperparameters are set as the original papers suggested.

- **Gradient** (Simonyan et al., 2014). This method computes the gradient of the target class *w.r.t.* each pixel as its attribution score. We use the codes⁴ released by the authors of Grad-CAM.
- **DeepLIFT** (Shrikumar et al., 2017). This backpropagation-based method assigns an attribution score to each unit, to reflect its relative importance at the original neural network input to the reference input. We use the codes released in the SHAP library⁵, where a black image is set as the default reference and the “rescale rule” is used to calculate the attribution.
- **IG** (Sundararajan et al., 2017). Integrated Gradient (IG) computes the modified gradient of the class *w.r.t.* each input pixel as the attribution score. More specifically, the modified gradient is the average gradient while the input varies along a linear path from a reference input to the original input. We use the codes⁶, where a black image is as the reference and the path steps are 100.
- **LRP** (Bach et al., 2015). Layer Relevance Propagation (LRP) starts at the output layer, redistributes the prediction of interest as the relevance of units in the previous layer, and recursively propagates the relevance scores until the input layer is reached. We use the official implementation⁷, where the “ ϵ -rule” is adopted to establish the relevance of single pixels.

²<https://github.com/pytorch/examples/tree/master/mnist>.

³<https://pytorch.org/docs/stable/torchvision/index.html>.

⁴<https://github.com/jacobgil/pytorch-grad-cam/blob/master/gradcam.py>.

⁵<https://github.com/slundberg/shap>.

⁶<https://github.com/TianhongDai/integrated-gradient-pytorch>.

⁷<http://heatmapping.org/tutorial/>.

- **XRAI** (Kapishnikov et al., 2019). This method treats the gradients of the target prediction *w.r.t.* region-level features as the attribution scores. In particular, it coalesces smaller regions into larger segments based on the maximum gain of IG per region. We use the official implementation⁸.
- **EP** (Fong et al., 2019). Extremal Perturbations (EP) optimize a spatial perturbation mask with a fixed area and smooth boundary that maximally affects a CNN’s prediction. We use the official implementation⁹.
- **RISE** (Petsiuk et al., 2018). Randomized Input Sampling for Explanation (RISE) masks an image using random occlusions patterns and observes change of confidence scores. For masking generation, RISE first samples smaller binary masks and then upsample them to larger resolution using bilinear interpolation. We use the official implementation¹⁰.
- **Grad-CAM** (Selvaraju et al., 2017). This method uses the gradient of the prediction *w.r.t.* feature maps in the last convolutional layer as the attribution scores, and arranges the attribution to the input image as the saliency map. We use the codes¹¹ released by the authors.

A.3 EVALUATION METRICS

Going beyond the visual inspections, we qualitatively evaluate the explanations of attribution methods on the task of image classification. Specifically, two evaluation protocols, Classification Accuracy@PI and Softmax Ratio@PI, can be computed as follows:

1. We start with a black image $\bar{\mathbf{x}}$;
2. For each image \mathbf{x} with the ground truth label c in the validation set \mathcal{V} , we form a ranking list of pixels \mathbf{m} in descending order, based on the saliency map of an attribution method $\mathbf{S}^c(\mathbf{x})$;
3. We add top PI (percentage of insertion) significant pixels \mathbf{m}_{PI} to the black image and get a masked image $\mathbf{x} \odot \mathbf{m}_{PI}$, where \odot is the element-wise product;
4. We then feed the masked image into the CNN model and get the prediction \tilde{c} and distribution over the classes $\tilde{\mathbf{p}}$.
5. We calculate two protocols as:

$$\text{Classification Accuracy@PI} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \mathbb{I}(\tilde{c} = c), \quad (8)$$

where $\mathbb{I}(\tilde{c} = c)$ is the binary indicator to evaluate whether the masked image is accurately classified, in order to measure the quality of explanations at a coarse granularity; and

$$\text{Softmax Ratio@PI} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{x} \in \mathcal{V}} \frac{\tilde{p}_c}{p_c}, \quad (9)$$

where \tilde{p}_c and p_c are the probability that the classes of masked or original images are equal to the ground truth c , in order to measure the attribution performance at a finer granularity.

In our experiment, we set PI as $\{5\%, 10\%, 15\%, \dots, 90\%, 95\%, 100\%\}$ and monitor the changes of each attribution method *w.r.t.* the protocols. Thereafter, we calculate the area under curves (AUC) of these two protocols, termed **Accuracy-AUC** and **Softmax-AUC** respectively.

B VISUAL INSPECTIONS OF A-FMI ON DIFFERENT CNN MODELS

We analyze the generalization ability of A-FMI by interpreting different CNN architectures, including the VGG-16, VGG-19, ResNet-50, ResNet-152, and ResNeXt-101-32x8d models.

Figure 7 shows the results of these five CNN models. When multiple objects of interest appear in a single image, VGG-16 and VGG-19 tend to distinguish them separately and perform better

⁸<https://github.com/PAIR-code/saliency>.

⁹<https://github.com/facebookresearch/TorchRay>.

¹⁰<https://github.com/eclique/RISE>.

¹¹<https://github.com/jacobgil/pytorch-grad-cam/blob/master/gradcam.py>.

localization; meanwhile, ResNet-50, ResNet-152 and ResNeXt-101-32x8d produce a larger region to cover several objects. When it comes to the single-object image, VGG-16 and VGG-19 might focus on the significant characteristics of the objects (*e.g.*, heads of animals), while ResNet-50, ResNet-152, and ResNeXt-101-32x8d favor finding the whole body of the object.

C REFERENCE RELIABILITY

We consider the variants of A-FMI, DeepLIFT, and IG that use different references — a black image, a white image, and an image filled with random pixels to generate the saliency maps. The results in Figure 8 correspond to the VGG-19 network.

Regardless of references, we observe that DeepLIFT and IG create grainy saliency maps and focus more attributions on the background or irrelevant objects (*e.g.*, the white snow background in the 1st column; the branch and grass highlighted in the 2nd column; the black background in the 3rd column; the person in the 4th column;) than within the objects of interest.

D OVERALL PERFORMANCE

We show visualizations of all methods with a percentage of important pixels insertion that vary from 5% to 50% in VGG-19. We compare A-FMI to popular attribution methods at three granular levels: pixel-wise (Gradient, DeepLIFT, IG, LRP), region-level (XRAI, EP, RISE), and feature map-based (Grad-CAM). Figures 10-14 clearly show that, compared with other popular attribution methods, A-FMI provides better visual explanations, in terms of more accurate and targeted localization of objects.

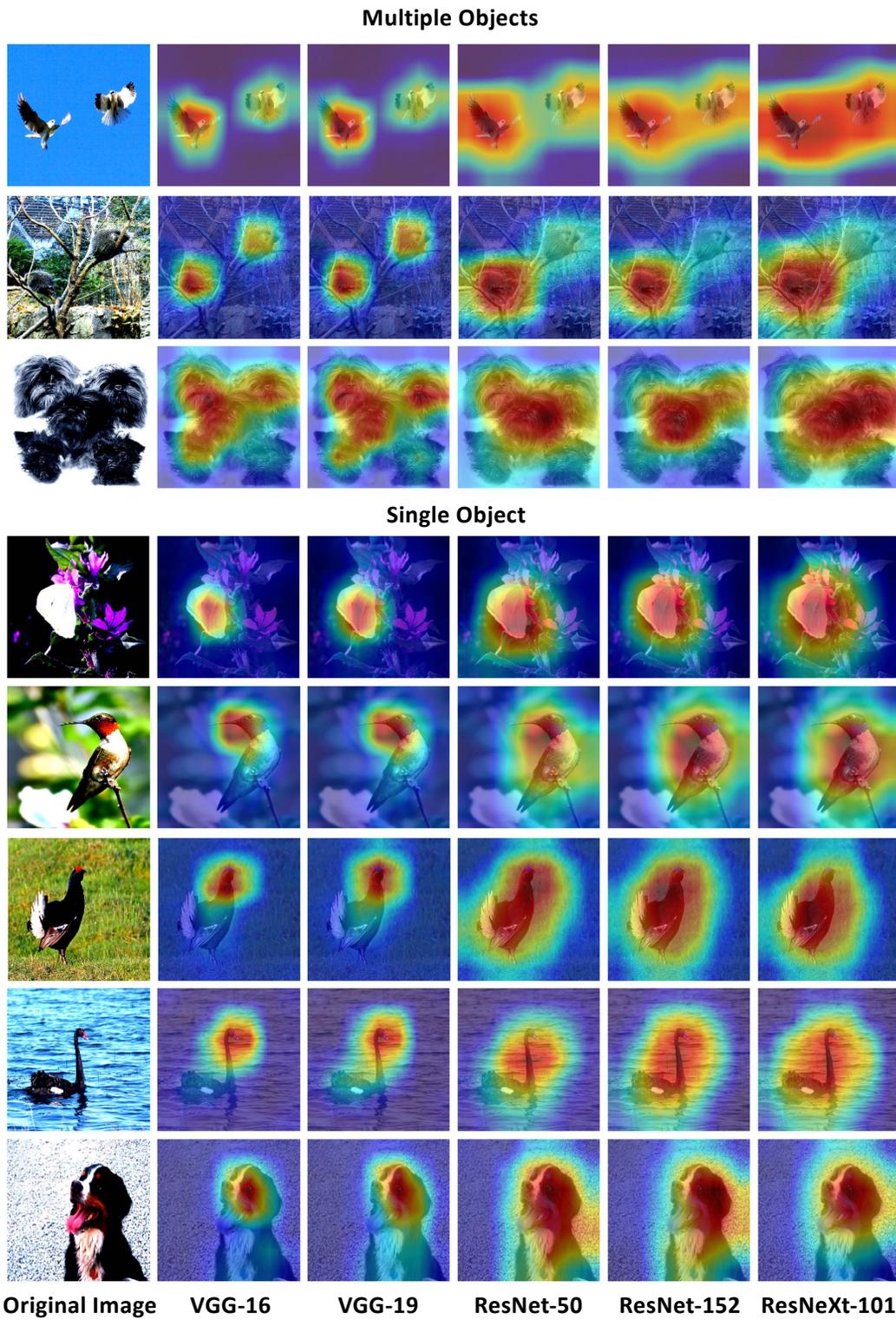


Figure 7: Visual Inspections of A-FMI on the VGG-16, VGG-19, ResNet-50, ResNet-152, and ResNeXt-101-32x8d models.

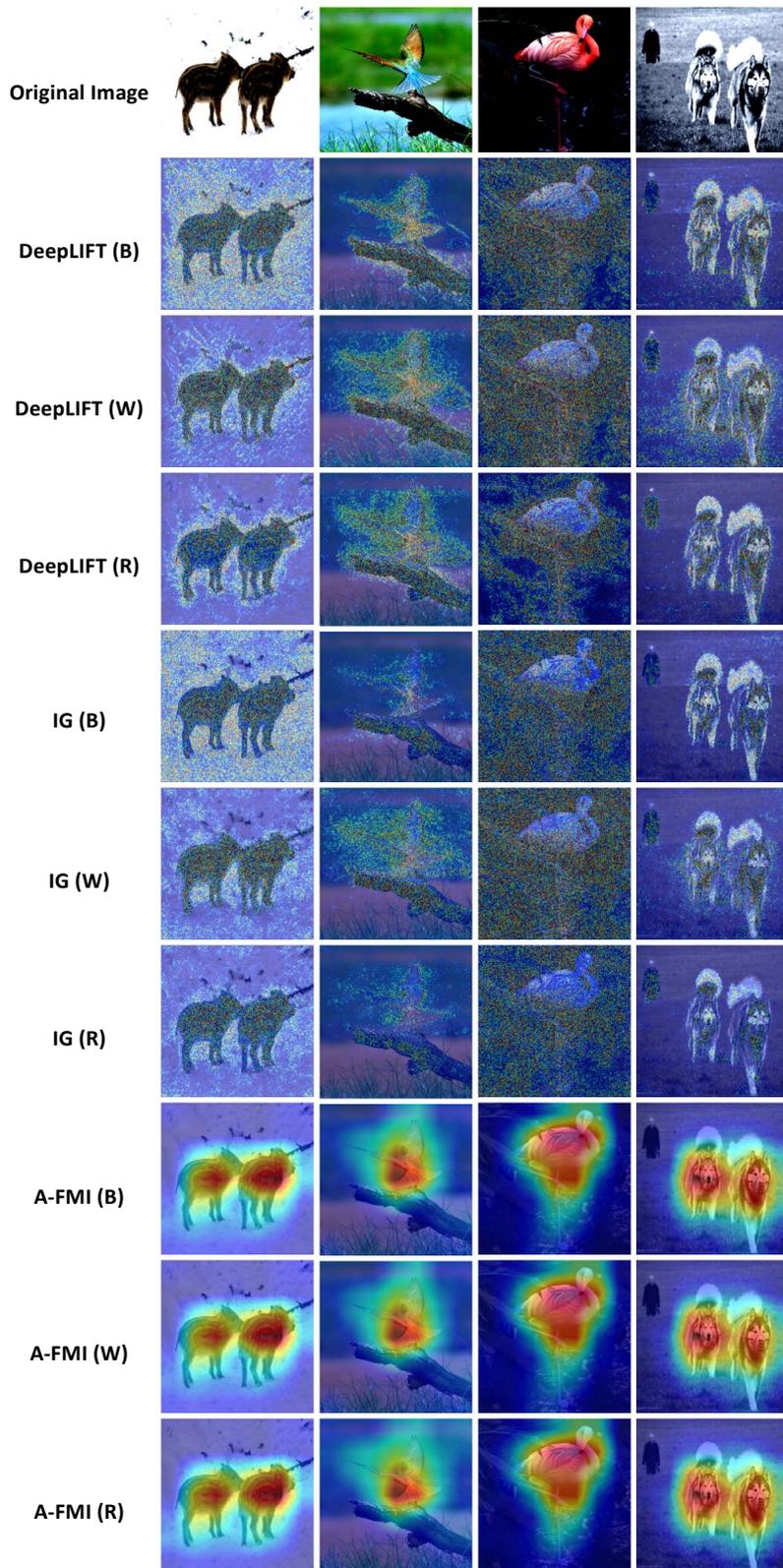
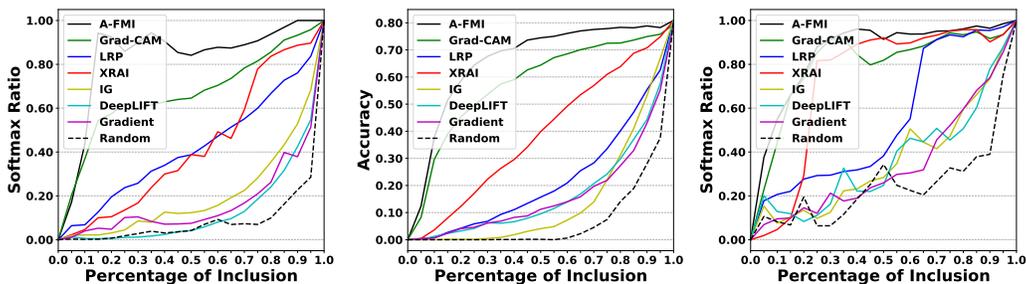


Figure 8: **Reference reliability.** A-FMI is reliable for the different choices of reference, while the explanations of DeepLIFT and IG rely heavily on the choice of references. Wherein, B, W, and R stand for black, white, and random reference respectively.



(a) Softmax Ratio curves on VGG19

(b) Accuracy & Softmax Ratio curves on VGG16

Figure 9: **Predictive Performance.** Accuracy curves and Softmax ratio curves of various attribution methods *w.r.t.* different percentage of important pixels insertion. Best viewed in color.

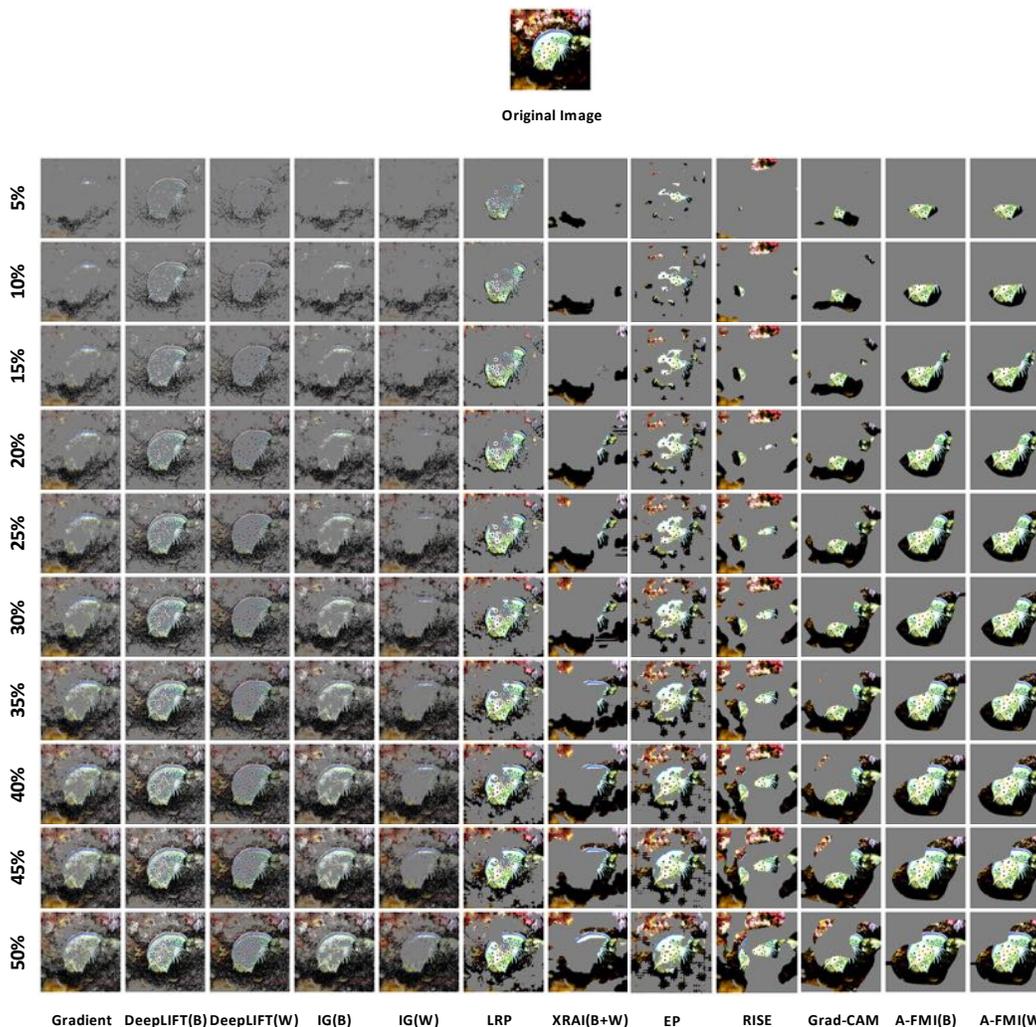


Figure 10: **Visual Inspections.** A visual comparison of all methods, where the percentage of important pixel insertion varies from 5% to 50%. Gradient, DeepLIFT and IG tend to produce grainy images. LRP, XRAI and Grad-CAM might choose disconnected areas. A-FMI focuses more on the objects of interest.

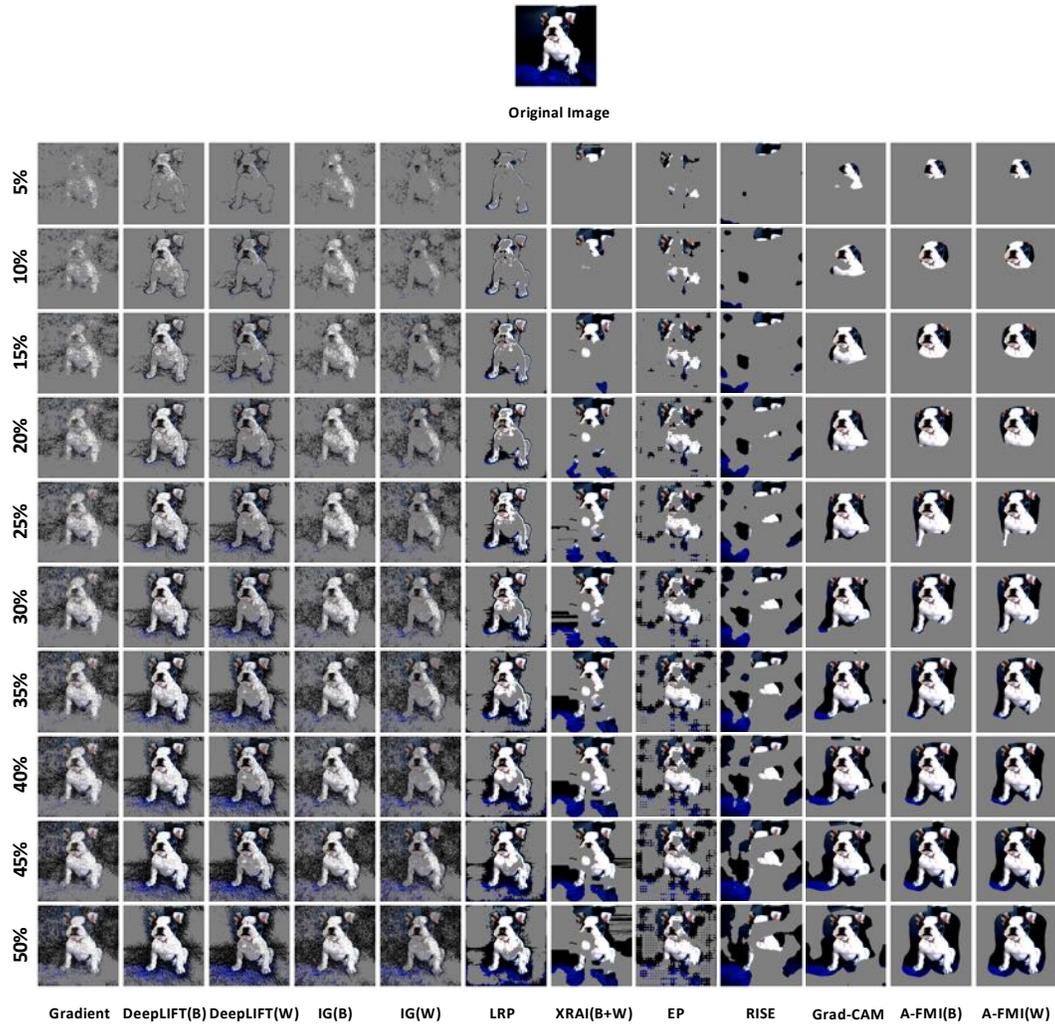


Figure 11: **Visual Inspections.** A visual comparison of all methods, where the percentage of important pixel insertion varies from 5% to 50%. Gradient, DeepLIFT and IG tend to produce grainy images. LRP, XRAI and Grad-CAM might choose disconnected areas. A-FMI focuses more on the objects of interest.

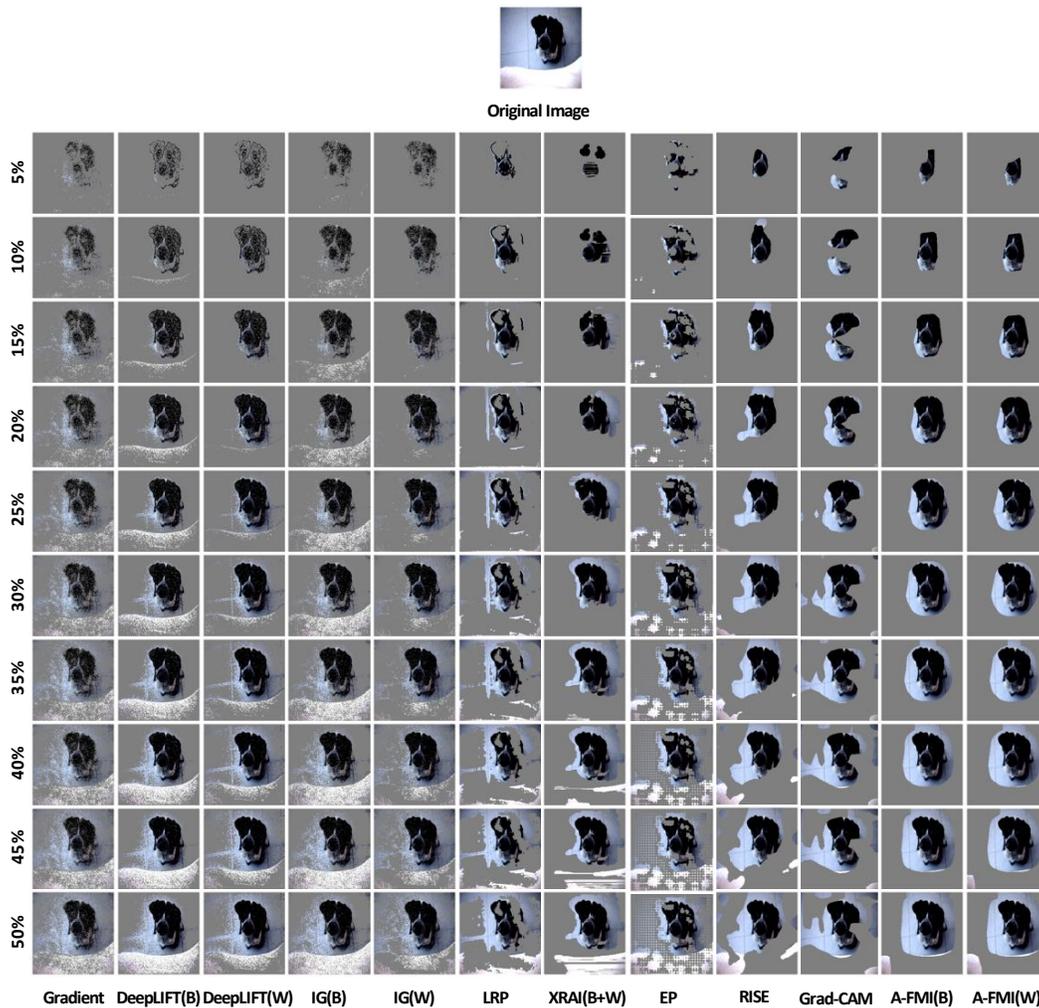


Figure 12: **Visual Inspections.** A visual comparison of all methods, where the percentage of important pixel insertion varies from 5% to 50%. Gradient, DeepLIFT and IG tend to produce grainy images. LRP, XRAI and Grad-CAM might choose disconnected areas. A-FMI focuses more on the objects of interest.

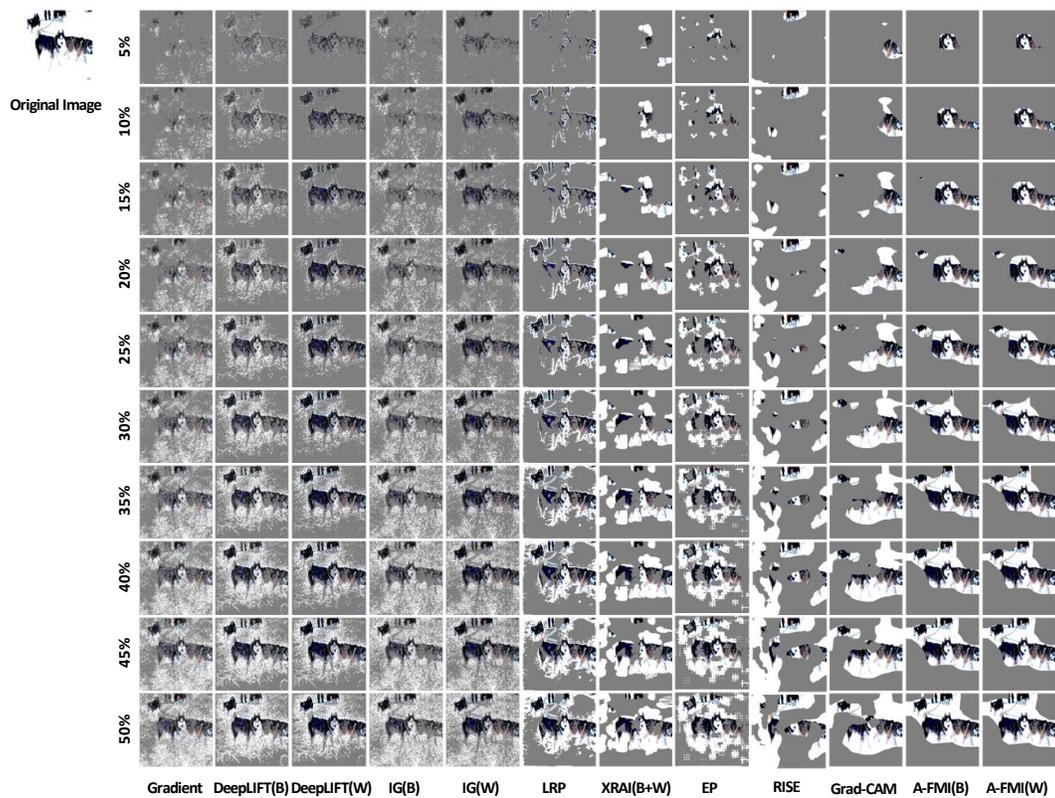


Figure 13: **Visual Inspections.** A visual comparison of all methods, where the percentage of important pixel insertion varies from 5% to 50%. Gradient, DeepLIFT and IG tend to produce grainy images. LRP, XRAI and Grad-CAM might choose disconnected areas. A-FMI focuses more on the objects of interest.

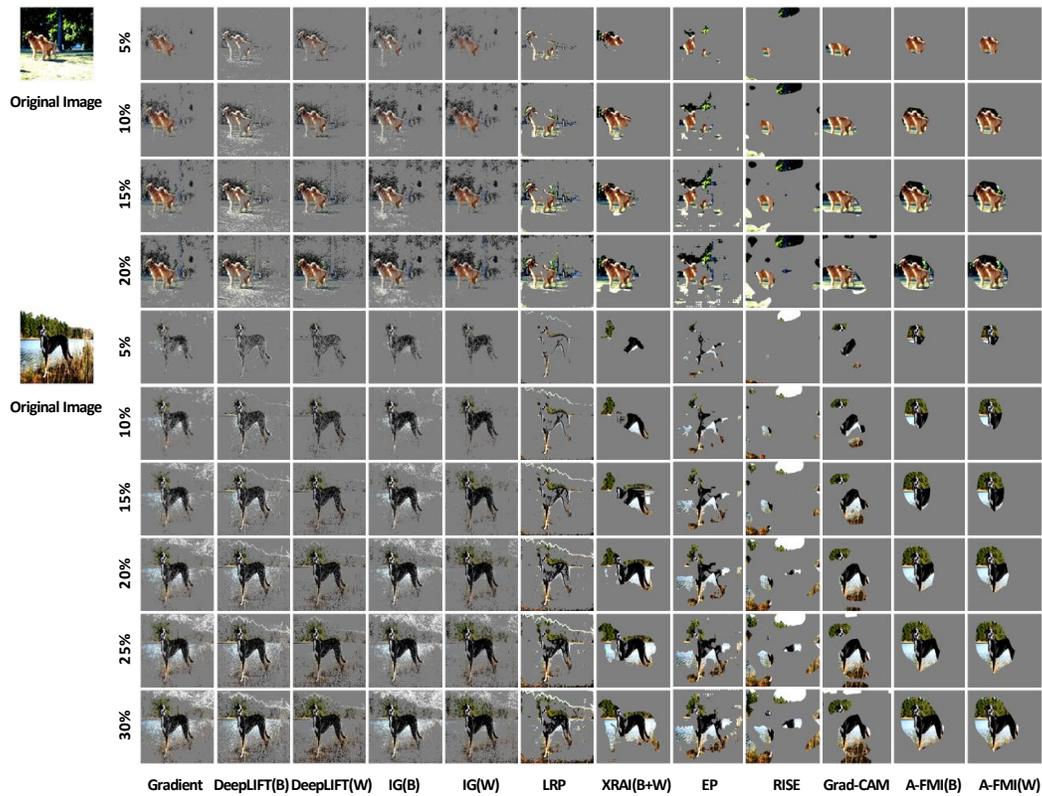


Figure 14: **Visual Inspections.** A visual comparison of all methods, where the percentage of important pixel insertion varies from 5% to 50%. Gradient, DeepLIFT and IG tend to produce grainy images. LRP, XRAI and Grad-CAM might choose disconnected areas. A-FMI focuses more on the objects of interest.