FovEx: Human-inspired Explanations for Vision Transformers and Convolutional Neural Networks

Mahadev Prasad Panda
Department AIBE,
FAU Erlangen-Nürnberg,
Erlangen, Germany
mahadev.prasad.panda@fau.de

Matteo Tiezzi
PAVIS,
Istituto Italiano di Tecnologia (IIT),
Genova, Italy
matteo.tiezzi@iit.it

Martina Vilas
Ernst Strüngmann Institute
for Neuroscience,
Frankfurt, Germany
martinagonzalezvilas@gmail.com

Gemma Roig
Goethe-Universität Frankfurt am Main
Frankfurt, Germany
roig@cs.uni-frankfurt.de

Bjoern M. Eskofier
Department AIBE,
FAU Erlangen-Nürnberg
Erlangen, Germany and
Institute of AI for Health,
Helmholtz Zentrum München,
Munich, Germany
bjoern.eskofier@fau.de

Dario Zanca
Department AIBE
FAU Erlangen-Nürnberg
Erlangen, Germany
dario.zanca@fau.de

Abstract

Explainability in artificial intelligence (XAI) remains a crucial aspect for fostering trust and understanding in machine learning models. Current visual explanation techniques, such as gradient-based or class-activation-based methods, often exhibit a strong dependence on specific model architectures. Conversely, perturbation-based methods, despite being model-agnostic, are computationally expensive as they require evaluating models on a large number of forward passes. In this work, we introduce Foveation-based Explanations (FovEx), a novel XAI method inspired by human vision. FovEx seamlessly integrates biologically inspired perturbations by iteratively creating foveated renderings of the image and combines them with gradient-based visual explorations to determine locations of interest efficiently. These locations are selected to maximize the performance of the model to be explained with respect to the downstream task and then combined to generate an attribution map. We provide a thorough evaluation with qualitative and quantitative assessments on established benchmarks. Our method achieves state-of-the-art performance on both transformers (on 4 out of 5 metrics) and convolutional models (on 3 out of 5 metrics), demonstrating its versatility among various architectures. Furthermore, we show the alignment between the explanation map produced by FovEx and human gaze patterns (+14% in NSS compared to RISE, +203% in NSS compared to GradCAM). This comparison enhances our confidence in FovEx's ability to close the interpretation gap between humans and machines.

Keywords Foveation-based Explanation · Human-inspired · Explainable Artificial Intelligence

1 Introduction

In recent years, deep learning has made remarkable strides in revolutionizing computer vision, particularly in safety-critical domains such as medical imaging [1, 2, 3], autonomous driving [4, 5], industrial automation [6, 7, 8], or security and surveillance [9, 10, 11]. However, as these models become increasingly more complex, the lack of understanding of their decision-making processes poses significant challenges[12]. To address these challenges, there is a growing need for XAI methods aiming at ensuring transparency and interpretability to the black-box nature of deep learning models.

While a variety of explanation methods have been developed for vision models [13, 14, 15, 16], these approaches are often tailored to specific architectures and lack universality. GradCAM and its derivations [13, 14] have been originally described as effective class-specific XAI methods to compute gradient-weighted feature maps from the last layer of convolutional architectures, highlighting relevant regions in the input. Although the GradCAM method can be extended to vision transformers (through reshaping the feature maps and gradients from the deepest layers), the performance of this approach is adversely affected by certain architectural attributes of vision transformers, such as skip connections, non-local self-attention mechanisms, and unstable gradients [17]. On the other side, XAI methods for vision transformers [18, 15, 16] are often tailored to transformer-specific characteristics, such as attention weights or class tokens, making their application to convolutional-based models unfeasible. Therefore, there is a pressing need for *model-agnostic* XAI methods, i.e., an approach that can work on any model architecture without the need for changes or adaptations. This can ensure comparability in explanations across different architectures and ensure more reliable interpretations of deep neural networks.

While current XAI methods provide insights into model decision-making processes, their quality often falls short when it comes to human understanding as they lack contextual aspects that make such explanations understandable to humans [19, 20]. Incorporating human-inspired constraints into XAI frameworks can enhance the quality of explanations and make them more aligned with human perception. One such fundamental aspect is foveated vision: humans' highest visual acuity occurs at the center of the visual field (fovea), while peripheral vision has a lower resolution, underlying the way humans prioritize details in a specific area. Recent work [21, 22] has demonstrated the advantages of incorporating such constraints into vision models.

Deza *et al*. [23] show increased i.i.d. generalization as a computational consequence emerging with foveated processing. Location-dependent computation based on foveation have demonstrated efficiency and avoidance of spurious correlation from data, both for convolutional [24] and visual transformer [25] models. Foveation priors are effective in generating visual scanpath [26]. These results together demonstrate how introducing biological constraints in artificial neural networks both increases alignment with the human counterpart and fosters model performances. We believe that the aforementioned concepts and intuitions can open a promising novel avenue in the field of explainability.

In this paper, we address the challenges associated with current explainability approaches and propose a novel method, Foveation-based Explanations (FovEx), that is based on insights inspired by the biology of human vision. The final goal of FovEx is to determine explanations, in the form of attribution maps, on the output of a backbone model processing a given input pattern. To do so, input samples undergo a differentiable transformation mimicking human-foveated vision. Then, gradient information is leveraged to enable an iterative and post-hoc human-like exploration of the input image, to determine relevant image regions.

Such relevant regions are combined appropriately to generate attribution maps. The FovEx-generated explanation maps, compared in Figure 1 against state-of-the-art methods (further details in the remainder of the paper), achieve state-of-the-art performance in common XAI metrics and have a better alignment with human gaze, as per experimental results.

To summarize, we delineate our contributions as follows:

- We introduce FovEx, a novel explanation method for DNNs that incorporates the biological constraint of human-foveated vision. FovEx extracts human-aligned visual explanations of the predictions in a post-hoc fashion, i.e., without introducing any architectural modification to the underlying predictor.
- We demonstrate the effectiveness of FovEx through qualitative and quantitative evaluations for both convolutional and transformer-based models. We compare our approach to class-activation-based and perturbation-based XAI methods and demonstrate state-of-the-art performance for different model architectures.
- We show that FovEx explanations enhance human interpretability via a quantitative investigation of the correlation between human gaze patterns and the explanation maps of DNNs generated by FovEx.

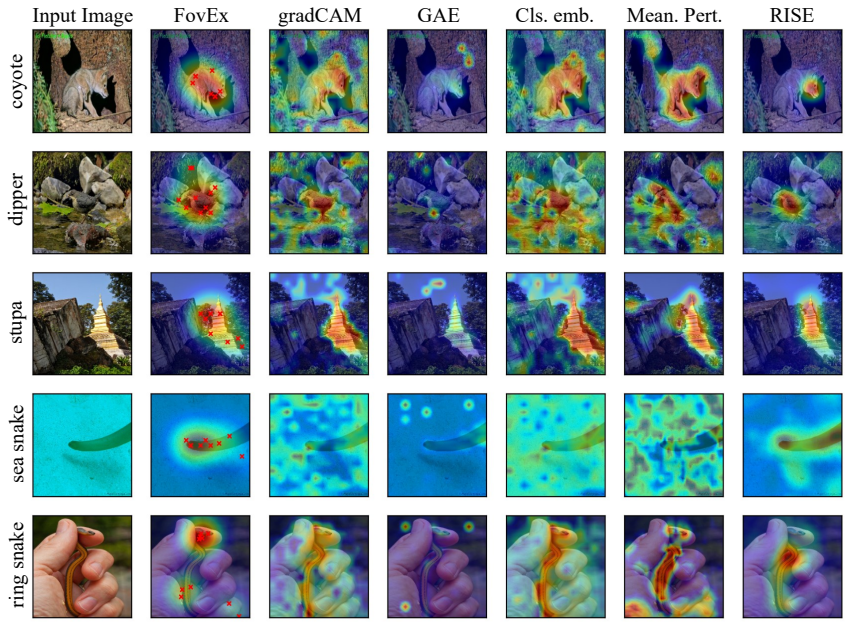

Figure 1: ViT-B/16 attribution maps. Explanation maps generated by FovEx (second column) and competitors for the ViT-B/16 predictor. Red crosses on explanation maps in the FovEx column denote fixation locations.

2 Related Work

A variety of local post-hoc XAI methods exist for supervised image classification models [27]. These techniques can be broadly categorized into (*i*) backpropagation or gradient-based methods, (*ii*) Class Activation Map (CAM)-based methods, and (*iii*) perturbation-based methods [28, 29]. In this section, we first introduce the main literature for these three categories, and then provide concise insights into specific methods tailored for the interpretability of transformer-based models.

Gradient-based Explanation Methods. To generate explanations, gradient-based XAI approaches utilize the gradient of a pre-trained black-box model's output with respect to input features, i.e., image pixels [30]. The seminal work by Simonyan *et al*. [31] constructs saliency maps by computing gradients of the non-linear class score function with respect to the input image. However, gradients frequently result in noisy visualizations. To address this issue, [32] refines the explanation map by averaging over multiple saliency maps for a single input image. These saliency maps correspond to noisy duplicates of the input image, built by introducing random Gaussian noise with a zero mean into the original image. Layer-wise Relevance Propagation (LRP) [33] generates an explanation map by propagating fixed predefined decomposition rules from the output layer of a black-box model to the input layer. Softmax Gradient Layer-wise Relevance Propagation (SGLRP) [34] utilizes the gradient of the softmax function to propagate decomposition rules, aiming to address the class-agnostic nature observed in vanilla LRP. Unlike gradient-based methods, our proposed approach uses gradient information exclusively to enable a human-like exploration of the input image to generate locations of interest.

CAM-based Explanation Methods. A class activation map (CAM) [35] reveals the important areas in an image that a convolutional neural network relies on to recognize a particular class. CAM leverages the Global Average Pooling (GAP) layer and the top-most fully connected layer of convolution-based classification networks. Even if CAM produces class-discriminating saliency maps, it is fully dependant on architectural families and constraints. GradCAM [13] extends CAM by incorporating gradient information. It computes the gradient of the predicted class score with respect to the feature maps of the last convolutional layer. These gradients act as the weights of each feature map for the target class. GradCAM++ [14] builds on GradCAM by employing pixel-wise weights rather than a single weight for a forward feature map of the final convolution layer. This enhancement enables GradCAM++ to preserve multiple instances of similar objects in the final explanation map. To further improve on the gradient-based CAM techniques, [36] uses class score as weights and [37] employs relevance score as weights in the explanation generation process. In contrast to CAM-based methods, our method is architecture agnostic, making it applicable to convolution and transformer-based models without any modifications.

Perturbation-based Explanation Methods. In the context of supervised image classification, perturbation-based explanation methods involve techniques that generate explanations by directly manipulating the pixels in the input image and observing the resultant changes in the output of the black-box model [38]. Various existing explanation techniques belong to this category [39, 40, 41, 28]. Perturbation-based explanation methods are particularly versatile, holding the potential for application across various model architectures. However, these approaches exhibit high computational complexity as they rely on the computation of a large number of perturbations and model inference steps. Unlike common perturbation methods, our approach strategically determines optimal focus locations using gradient information, making it substantially more computationally efficient, see Section 4.3.

Explaining Transformers. While there is a variety of explanation methods available for convolution-based models, the range of methods for transformers is relatively limited. Chefer *et al*. [15] contribute to this domain by generating explanation maps for transformers with a method that combines LRP and gradient-based approaches.[16] extends the application of the method proposed in [15] to provide explanations for any type of transformer model. Additionally, Vilas *et al*. [42] propose an approach that quantifies how different regions of an image can contribute to producing a class representation in intermediate layers, using attention and gradient-based information. However, it is important to note that these methods cannot be seamlessly employed in a plug-and-play manner without making adjustments to pre-existing model implementations. Conversely, in this paper, we propose a novel method that can be applied to both convolution-based and transformer-based architectures without altering the neural architectures for computing the explanation.

3 Foveation-based Explanation: FovEx

Let us consider a black-box predictor $b(\cdot|\theta)$ defined for classification problems, that, without any loss in generality, we assume to be a neural network with learnable parameters θ. The predictor is a function $b : \mathcal{X} \in \mathbb{R}^i \mapsto \mathcal{Y}$, that maps data instances x from the input space $\mathcal{X} \in \mathbb{R}^i$ to the prediction y in the target space $\mathcal{Y}$, containing the different labels to which the input data pattern can belong. We denote with $y = b(x|\theta)$ the prediction y yielded by the predictor on the input pattern x. We denote with θ_D the case in which the learnable parameters θ have been tuned on a training dataset D.

The goal of FovEx is to extract a human-understandable visual explanation of the decisions taken by the predictor b, in a post-hoc fashion, i.e., without introducing any architectural modification to the underlying predictor b. Formally, FovEx is a function

$$E = \text{FovEx}(b, x) \tag{1}$$

where the output E denotes the attribution map for the predictor b associated to the input x. FovEx in turn consists of three fundamental operations, i.e., (i) a differentiable foveation mechanism, (ii) a gradient-based attention mechanism, and (iii) an attribution map generation process. A schematic illustration of the method is given in Figure 2. In the following, we give a formal definition of each operation.

Differentiable Foveation. Biological foveated vision is characterized by a central area of fine-grained processing (i.e., the fovea) and a coarser peripheral area. We draw inspiration from Schwinn *et al*. [26] to design a differentiable foveation mechanism. Let x be an input image and f_t be the current coordinates of the focus of attention at time t. First, we define a coarse version of x, denoted by $\bar{x}$, by convolving it with a Gaussian kernel, i.e.,

$$\bar{x} = x * \mathcal{G}(0, \sigma_b^2) \tag{2}$$

where $*$ represents the convolution operation, and $\mathcal{G}(0, \sigma_b^2)$ denotes the Gaussian kernel with a mean of 0 and a standard deviation of σ_b. The parameter σ_b controls the amount of blurring in the periphery of the image. From a biological standpoint, $\bar{x}$ can be regarded as the fundamental information obtained by peripheral vision within the initial milliseconds of stimulus presentation. A foveated input image x_Φ is obtained as a weighted sum of the original input x and the coarse version of the input $\bar{x}$, i.e.,

$$x_\Phi = \Phi\left(x, f_t\right) = \mathcal{W}(t) \cdot x + (1 - \mathcal{W}(t)) \cdot \bar{x} \tag{3}$$

where $\cdot$ denotes the pixel-wise multiplication. In equation 3, $\mathcal{W}(t) = \mathcal{G}\left(f_t, \sigma_f^2\right)$ stands for the pixel-wise weighting factor defined as the Gaussian blob $\mathcal{G}\left(f_t, \sigma_f^2\right)$ with a mean of f_t and a standard deviation of σ_f, and $\Phi(\cdot)$ represents the foveation function. It is important to notice that, since all operations are differentiable, it makes the propagation of gradient information necessary for subsequent steps feasible. Additionally, the transformation described above introduces noise according to a foveal distribution $\mathcal{W}(t)$, perturbing the original input.

Gradient-based Attention Mechanism. The foveation mechanism allows for sequential exploration of the given input image. The next location of interest (i.e., the next fixation point) will depend on the state s_t generated by all previous

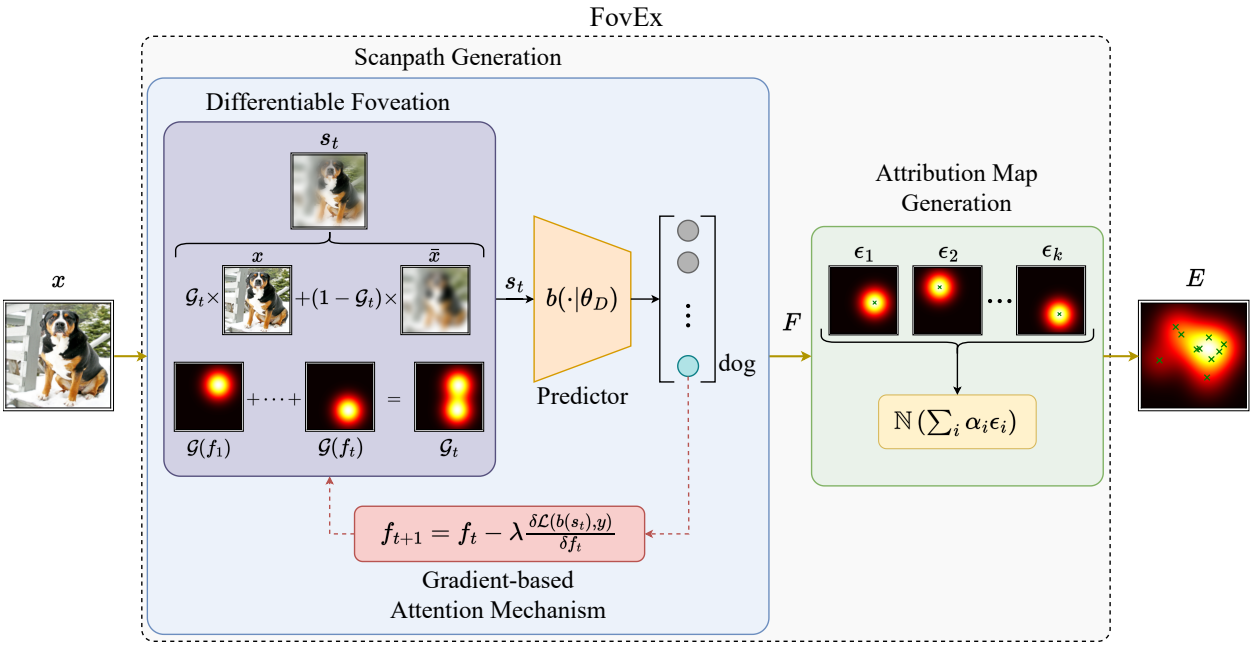

Figure 2: The proposed FovEx. Given an input image x, FovEx produces an attribution map E for a predictor b trained on a dataset D. The image x undergoes a differentiable foveation process yielding a transformed input (s_t) for the predictor b. The loss function $\mathcal{L}$, computed based on predicted class scores and ground truth y, is exploited by an attention mechanism for the generation of a sequence of fixations $(f_1, f_2, \cdots, f_t)$ (referred to as scanpath F). The resultant scanpath F is employed to build a weighted linear combination of individual saliency maps ϵ_i associated with each fixation point f_i, leading to the final attribution map E.

fixation locations, which can be regarded as the system's memory. The s_t is obtained by cumulating Gaussian blobs to gradually expand the region of good visual fidelity after each fixation point, i.e.,

$$s_t = s(x, f_t) = \mathcal{G}_t \cdot x + (\mathbb{1} - \mathcal{G}_t) \cdot \bar{x} \tag{4}$$

where $\mathcal{G}_t = \sum_{j=0}^{t} \beta^j \mathcal{G}_{t-j} \left(f_j, \sigma_f^2\right)$ symbolizes the cumulative Gaussian blob, whereas $\mathbb{1}$ refers to a square matrix of ones. The forgetting factor $0 \leq \beta \leq 1$ regulates how much information is retained from previous fixations.

Let $\mathcal{L}\left(b(s_t), y\right)$ be the loss function at time t, e.g., calculated as the distance between the output predicted by b for the current state s_t and the target class y_t. The next fixation locations are dynamically adjusted to minimize the loss function $\mathcal{L}$ with respect to the current fixation location f_t, i.e.,

$$f_{t+1} = f_t - \lambda \frac{\delta \mathcal{L}}{\delta f_t} \tag{5}$$

where the hyperparameter λ determines the step size at each iteration during optimization. The optimization technique iterates until a specified number of optimization steps (OS) have been performed, ensuring successful convergence. The influence of OS is discussed in the ablation studies, in Section 4.4. Random restarts (RR) are implemented when optimization fails to yield improvements in the loss function after a specified number of steps OS, serving as a strategy to escape local minima. New fixations can be generated for an arbitrary number N of steps, resulting in a sequence of N fixations points, also called a *scanpath*

$$F = (f_1, ..., f_N) \tag{6}$$

Attribution Map Generation. At this point, we have generated a sequence of N fixation points based on an input and a task model. Each fixation point f_i can be associated with a saliency map ϵ_i, describing a 2D Gaussian distribution with a mean at f_i and a standard deviation σ_ϵ, where σ_ϵ is set to match the standard deviation of the Gaussian blob (σ_f). The final attribution map E, functioning as an explanation for the predictor b, is obtained as a weighted linear

combination of the individual saliency maps associated with each fixation point, i.e.,

$$E = \mathcal{N}\left(\sum_{i=1}^{k} \alpha_i \epsilon_i\right) \tag{7}$$

In Equation 7, α_i denotes the weighting factor for saliency map ϵ_i and $\mathcal{N}(\cdot)$ represents min-max normalization. The weights α_i determine the contribution of each fixation to the final saliency map. In our experiments, we set $\alpha_i = 1$, $\forall I \in \{1, ..., N\}$, as different weighting schemes did not improve the quality of explanations on a validation set.

4 Experiments

We conducted a comprehensive set of experiments to compare the performances of the proposed FovEx against state-of-the-art (SOTA) models and to showcase its ability to be agnostic to the architecture of the predictor b. We assessed FovEx's performances in different scenarios, ranging from qualitative inspections and quantitative assessments in the common testbed of ImageNet-1K validation set [43] to the correlation analysis of the generated attribution maps to human gaze. Additionally, we compared FovEx's computational complexity against SOTA methods and performed in-depth model ablation studies. All our experiments were performed in a Linux environment, using an NVIDIA RTX 3080 GPU, and the implementation code can be found at `https://github.com/mahadev1995/FovEx`.

4.1 ImageNet-1K

Setup & Data. We selected two representative classification models as the predictor b to be explained. In particular, we focused on a ResNet-50[1] [44] and a Vision Transformer (ViT-B/16)[2] [45], that have been recently classified as foundation models [46]. The ResNet-50 model is pre-trained on the ImageNet-1K [43] dataset, while the ViT-B/16 model is pre-trained on ImageNet-21K [47] dataset and fine-tuned on ImageNet-1K [43] dataset. We assessed the model performances on a subset of 5000 images from the ImageNet-1K [43] validation set, randomly sampled among the ones correctly classified by the predictor b in order to measure the contribution of the method exactly, as pointed out by recent literature [37]. The images are resized to a resolution of 224×224 pixels and normalized in the range of $(0, 1)$. Performance on additional models is presented in Appendix A.

Metrics. We report model performances focusing on the faithfulness and localization attributes of the explanation maps. Faithfulness measures the extent to which explanation maps correspond to the behavior of the black-box model. We report AVG. % DROP (lower is better) and AVG. % INCREASE introduced in [14] (higher is better), as well as the DELETE (lower is better) and INSERT metrics (higher is better) proposed by Petsiuk *et al*. [40]. The AVG. % DROP metric assesses the shift in confidence between two scenarios: one with the entire image as input and another with input limited to the regions highlighted by the explanation map. The AVG. % INCREASE metric quantifies instances across the dataset where the model's confidence rises when only the highlighted regions from the explanation map are considered. The DELETE metric is used to evaluate the decrease in predicted class probability when removing pixels with decreasing importance, according to the explanation map. On the other hand, the INSERT metric measures the increase in estimated likelihood when adding the essential pixels to the input, ordered from most to least important as indicated by the attribution map. Localization measures an explanation map's capability to focus on a specific region of interest. Although good performance on localization may not imply a good explanation, it can provide interesting insights nonetheless because explaining a model's localization decisions can help identify patterns in the data the model has learned [48]. We report the Energy-Based Pointing Game (EBPG) metric [36] (higher is better). The EBPG metric calculates the energy of attribution maps within a predefined bounding box of the target class.

Compared Models & Architecture Details. We compared the performances attained by FovEx against various SOTA XAI techniques. When considering the ViT-B/16 predictor, we report performances obtained by gradCAM [13], Meaningful Perturbation (Mean. Pert.) [39], RISE [40], GAE [16], and class embedding projection (Cls. Emb.) [42]. We remark that perturbation-based methods (Mean. Pert., RISE) have not been previously tested with Transformer architectures. In the case of a ResNet-50 predictor, we consider gradCAM, gradCAM++ [14], Mean. Pert., and RISE as competitors. In all the settings, we also report a baseline technique for sanity check, referred to as RandomCAM. In particular, randomCAM generates class activation maps for a randomly selected class regardless of the black-box model's prediction. For competitors, we utilized hyper-parameters following the respective original implementations. When considering FovEx, we adhere to the parameters suggested in Schwinn *et al*. [26], which are biologically plausible, as they are designed to mimic human vision. The effect of different choices for the hyperparameters is discussed in our ablation study discussed in Section 4.4.

[1] `https://pytorch.org/vision/stable/models.html`

[2] `https://github.com/google-research/vision_transformer`

Table 1: ViT-B/16 quantitative evaluation. Average metrics on the considered subset of ImageNet-1K validation dataset. The best-performing model is in bold, and the second-best is underlined.

Eval. Name	FovEx	grad CAM	GAE	Cls. Emb.	Mean. Pert.	RISE	random CAM
AVG. % DROP (↓)	**13.970**	40.057	86.207	34.862	29.753	<u>15.673</u>	80.714
AVG. % INCREASE (↑)	**30.389**	11.469	0.799	13.329	20.549	<u>22.189</u>	1.789
DELETE (↓)	0.240	<u>0.157</u>	0.172	**0.155**	0.200	0.158	0.395
INSERT (↑)	**0.840**	<u>0.818</u>	0.806	0.817	0.674	0.782	0.682
EBPG (↑)	**47.705**	41.667	39.812	39.350	40.646	<u>42.633</u>	35.708

Table 2: ResNet-50 quantitative evaluation. Average metrics on the considered subset of ImageNet-1K validation dataset. The best-performing model is in bold, and the second-best is underlined.

Eval. Name	FovEx	grad CAM	grad CAM++	Mean. Pert.	RISE	random CAM
AVG. % DROP (↓)	**11.780**	21.718	19.863	85.973	<u>11.885</u>	61.317
AVG. % INCREASE (↑)	**61.849**	43.669	45.069	4.700	<u>55.489</u>	16.729
DELETE (↓)	0.151	0.108	0.113	**0.082**	<u>0.100</u>	0.212
INSERT (↑)	**0.374**	0.368	0.361	0.280	<u>0.372</u>	0.287
EBPG (↑)	46.977	**48.658**	<u>47.412</u>	42.725	43.312	38.118

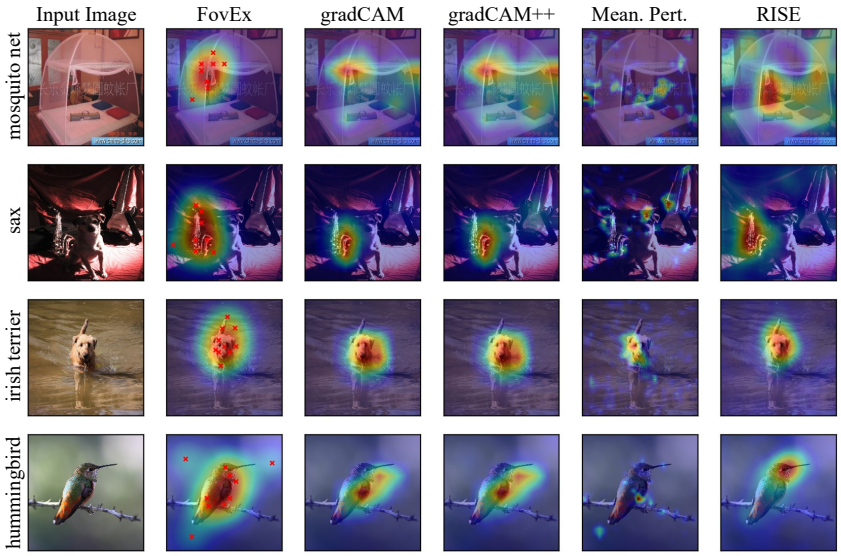

Figure 3: ResNet-50 attribution maps. Explanation maps generated by FovEx and competitors for the ResNet-50 model. Red crosses on explanation maps in the FovEx column denote fixation locations.

Quantitative Inspection. We report in Table 1 the results obtained for the ViT-B/16 predictor. FovEx outperforms all the competitors across all metrics, with the exception of DELETE, where the Cls. Emb. method excels. RISE secures the second-best position in AVG. % DROP, AVG. % INCREASE, and EBPG metrics, while gradCAM achieves the second-best performance in DELETE and INSERT. Similar conclusions can be drawn when we consider b to be a ResNet-50, as reported in Table 2. FovEx overcomes other methods in AVG. % DROP, AVG. % INCREASE, and INSERT, while delivering competitive results in DELETE and EBPG. We remark that while Mean. Pert. performs better than FovEx when considering DELETE, it exhibits worse overall performances in all the other considered metrics. GradCAM is the top performer when considering EBPG, followed by gradCAM++. RISE secures the second position in all other metrics, except for EBPG.

These results showcase better faithfulness and localization capabilities of the proposed FovEx as compared to other methods, independent of the underlying model architecture. The poorer performances of FovEx on DELETE can

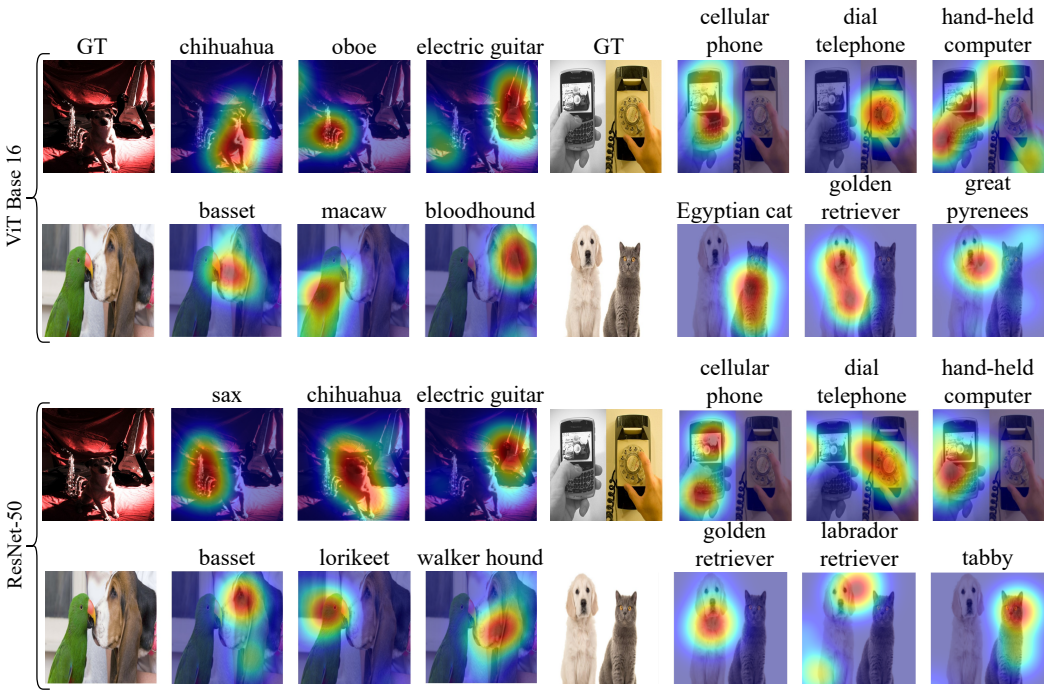

Figure 4: Class-discriminative evaluation. Class-specific attribution maps generated from FovEx, using ViT-B/16 (top) and ResNet-50 (bottom) as the predictor.

be attributed to the performed optimization scheme. Indeed, FovEx operates on a preservation ideology, where the optimization scheme aims to identify regions in the image that, when retained, allow the model to maintain its performance in correctly classifying input images. This design predisposes FovEx towards excelling in the INSERT metric, which better aligns with FovEx optimization objective, but it exposes to downfalls when considering DELETE. This is because the optimization process inherently focuses on finding the most critical regions for maintaining classification accuracy, rather than those whose removal would most degrade performance, leading to penalization in the DELETE metric.

Qualitative Inspection. To better assess the explanations provided by the proposed FovEx, we conducted a qualitative evaluation comprising both a (*i*) visual comparison concerning the considered competitors and (*ii*) in-depth investigation regarding class-specific explanations. Figure 1 illustrates a comparison of attribution maps generated on several (rows) input samples (first column) by the proposed FovEx (second column) and the other considered competitors (subsequent columns), in the case of the ViT-B/16 predictor. FovEx generates explanation maps that are spread over the whole object of interest without any blemish or stray spot, such as the ones produced by gradCAM, GAE, and Cls. Emb. An explanation of such phenomenon is given by [49], demonstrating the presence of artifacts in the feature map of models, corresponding to high-norm tokens appearing during inference in unimportant background areas of images, repurposed for internal computations. Indeed, we remark how the issue is solved when using FovEx as the explanation method. Similarly, RISE yields noise-free explanation maps, but we remark that its computational time is significantly higher compared to FovEx (see Section 4.3).

To better assess the quality of attribution maps generated by FovEx when dealing with convolution-based architectures, we report in Figure 3 a visual comparison of attribution maps generated by FovEx (second column) and other SOTA methods for the ResNet-50 predictor. More visualizations of attribution maps for different models are given in Appendix C. Also in this setting, FovEx is capable of generating attribution maps that are focused on the main object class contained in the processed image. The attribution map correctly spans over the object of interest, thanks to the exploration carried on by the attention scanpath, differently from competitors (RISE, gradCAM) that tend to attribute the class prediction to certain object parts (bird head).

Additionally, we tested FovEx's ability to localize class-discriminative information and capture fine-grained details. In particular, we selected images containing multiple annotated categories and we generated attribution maps for each of the available classes with FovEx. We report in Figure 4-top the attribution maps obtained with ViT-B/16 as the predictor when generating explanations for different categories (indicated in the column header). Attribution maps generated from

Table 3: Human-gaze correlation. Quantitative evaluation results from human gaze experiment. We report NSS and AUCJ metrics (higher is better, see the main text for further details.)

Eval. Name	FovEx	grad CAM	grad CAM++	Cls. Emb.	GAE	Mean. Pert.	RISE
NSS (↑)	**0.7160**	0.2357	0.0372	0.1231	0.4120	0.6197	0.6287
AUCJ (↑)	**0.7044**	0.5581	0.5094	0.5317	0.6875	0.6698	0.6400

FovEx can discern the desired categories, localizing the relevant category inside the image with a good fine-grained definition of the pixels belonging to the object. Similar conclusions can be drawn when considering ResNet-50 as the predictor, as depicted in Figure 4-bottom.

We report the FovEx performance when considering three additional predictors (b) such as ConvNeXt [50], ViT-B/16 and ViT-B/32 from `torchvision`[3] in the Appendix A. In summary, FovEx continues to outperform other SOTA methods in the majority of the evaluation metrics considered for the aforementioned models.

4.2 Human Gaze Correlation

Setup & Data. Generating attribution maps that are similar to the ones produced by human-gaze for image classification can enhance visual plausibility [51, 52]. Here, we investigate the correlation of attribution maps generated by FovEx and other competitors with the ones obtained from human gaze for image free-viewing task. The goal is to investigate the possibility of generating explanation maps correlated to the general human gaze while being faithful to a black-box model's decision-making. We exploit the MIT1003 dataset [53], composed of 1003 images that are complemented by corresponding human attention maps, collected in a controlled setting.

Metrics. To quantitatively assess the similarity between the attribution maps generated by the black-box model and the human attention map, we employ the Normalised Scanpath Similarity (NSS) and Area Under the ROC Curve Judd version (AUCJ) metrics [54]. The higher the value of such metrics, the better.

Compared Models & Architecture Details. We compare FovEx against the SOTA competitors described in the previous experiments. For this study, we focus on the ViT-B/16 predictor, pretrained on ImageNet-21K [47] and fine-tuned on ImageNet-1K [43].

Results. As reported in Table 3, FovEx outperforms other methods in both the considered metrics. The achieved performances are almost doubled with respect to gradCAM and gradCAM++. Remarkably, explanation methods that have been proposed for the tested architecture (GAE, Cls. Emb.) are outperformed. This result highlights the alignment between the explanation maps produced by FovEx and human gaze patterns during free-viewing of natural images and enhances our confidence in FovEx's ability to close the interpretation gap between humans and machines. Indeed, we remark that these results should be analyzed in conjunction with the ones about XAI metrics, summarised in Tables 2 and 1. Overall, to proposed FovEx is capable of coupling (i) the best alignment with human gaze patterns among the tested explanation methods, which we showed by investigating its correlation with human attention, together with (ii) extremely promising quantitative performances on several XAI metrics. This suggests that FovEx provides a more accurate and intuitive understanding of the model's decision-making process with respect to alternative methods, by allowing for a better comparison between human fixations and input localization importance of the model.

4.3 Efficiency Analysis

In this section, we investigate the time efficiency of the proposed FovEx compared with other SOTA approaches. We report the time necessary to produce an attribution map averaged over the considered 5000 images from the ImageNet-1K validation dataset. We focus our analysis on the ViT-B/16 predictor (a similar conclusion holds for the ResNet-50 architecture). In Figure 5-left we report the inference time (x-axis, seconds) against the AVG % DROP metric (y-axis) for all the considered methods. The closer to the left-bottom corner, the better. Figure 5-right depicts inference time (x-axis, seconds) against INSERT (y-axis). In this case, the left-top corner implies better performance. Overall, FovEx outperforms perturbation-based approaches (Mean. Pert. and RISE) in both performance and time efficiency. Conversely, gradient-based methods are more time efficient but are far from the performances achieved by FovEx in the considered metrics.

[3]Models from `torchvision` are trained on ImageNet-1K, while ViT-B/16 used in Section 4.1 is trained on ImageNet-21K and finetuned on ImageNet-1K. Source: `https://pytorch.org/vision/stable/models.html`

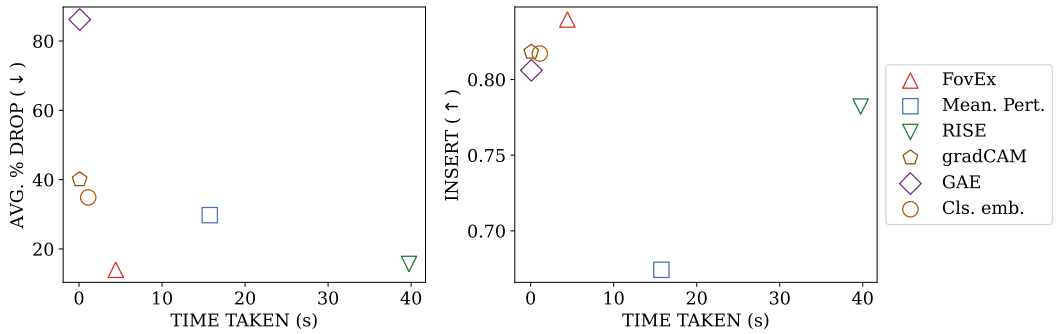

Figure 5: Efficiency analysis. Comparison of average attribution map generation time (x-axis, seconds) against model performance (y-axis), both when considering AVG % DROP (left) and INSERT metrics (right), for the ViT-B/16 model.

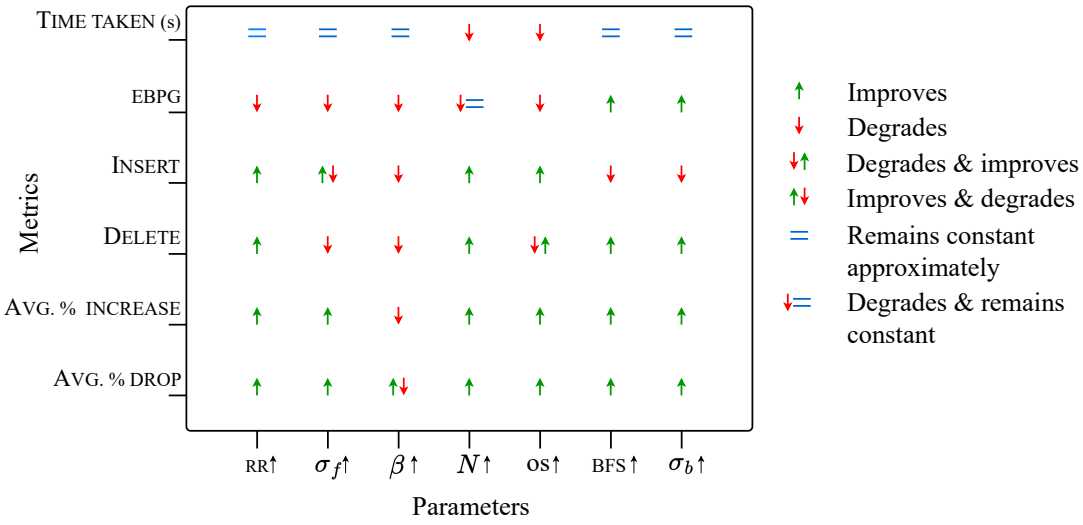

Figure 6: Summary of ablation study conducted to showcase the effect of various FovEx parameters. The ↑ near parameters (x axis) denotes an increase in their value; except for rr, it signifies the value shifts from False to True.

Table 4: Quantitative assessment. Average metrics on the considered subset of ImageNet-1K validation dataset for ViT-B/16 model for different RR values. Bold terms denote the best performance and underlined terms represent the second best performance.

Eval. Name	AVG. % DROP (↓)	AVG. % INCREASE (↑)	DELETE (↓)	INSERT (↑)	EBPG (↑)
RR = True	**13.973**	**30.389**	**0.240**	**0.840**	<u>47.705</u>
RR = False	<u>17.561</u>	<u>26.389</u>	<u>0.247</u>	<u>0.824</u>	**51.067**

4.4 Ablation studies

We undertook several ablation studies, which are delineated below. We conducted experiments to study the effect of various parameters such as random restart (RR), foveation sigma (σ_f), forgetting factor (β), scanpath length (N), optimization steps (OS), blur sigma (σ_b), and blur filter size (BFS). Figure 6 depicts the summarized impact of the parameters mentioned above on the performance of FovEx. We utilized the ViT-B/16 model and the corresponding subset of the ImageNet-1K validation dataset employed in Section 4.1 for the ablation studies. Default values for the parameters are given in Appendix B.

Random Restart (RR). The RR parameter helps to avoid local minima in the optimization scheme for finding the location of fixation f_t. When RR is set to True, the exploration starts at a completely random location. Figure 7 depicts

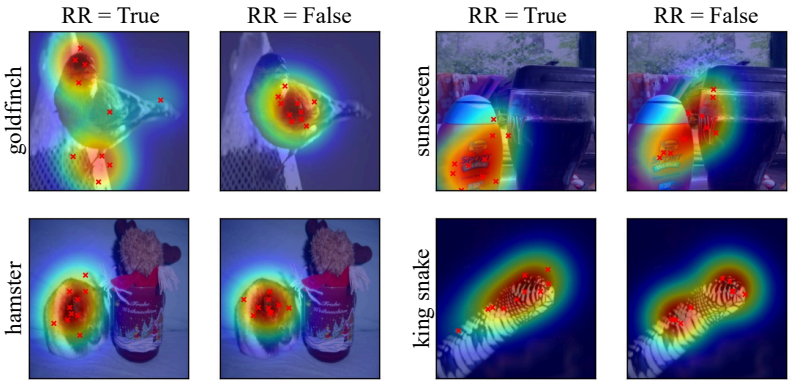

Figure 7: Qualitative assessment. Comparison of attribution maps generated using FovEx with different RR values.

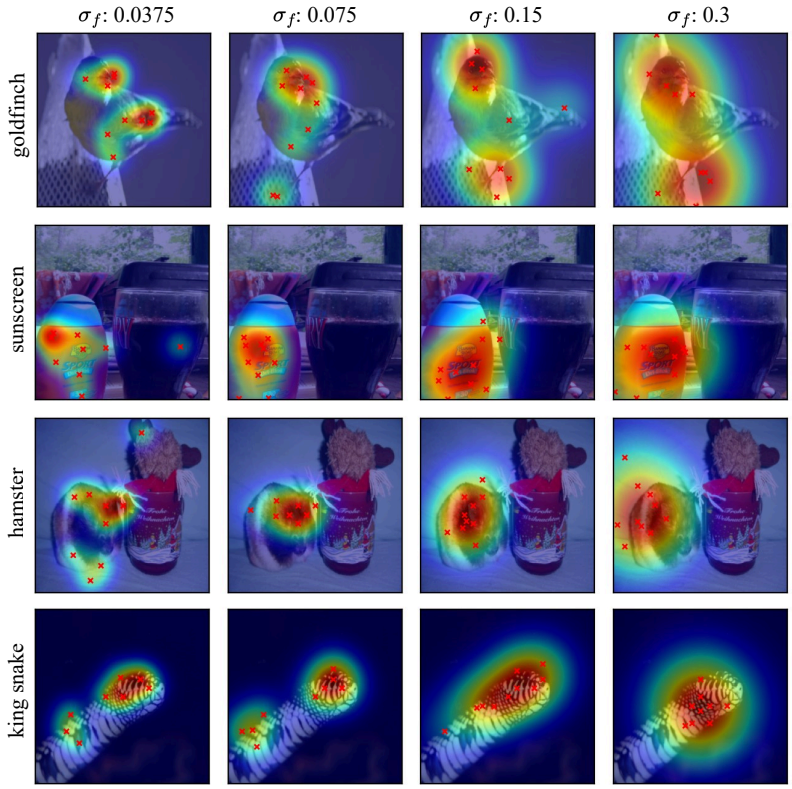

Figure 8: Qualitative assessment. Comparison of attribution maps generated using FovEx with different σ_f values.

the visual comparison of explanation maps generated using two different values of RR. On close inspection, it is evident that fixation locations are closer to each other when RR is set to False. Table 4 presents the comparison of quantitative evaluation for different configurations of RR. With RR set to True, we get better performance in every metric except for the EBPG metric.

Foveation Sigma (σ_f). The σ_f parameter controls the standard deviation of the area with higher visual acuity in the foveated input image. A higher value of σ_f results in a larger region with enhanced visual clarity. In our experiments, we create attribution maps with varying σ_f values- specifically, $\{0.3, 0.15, 0.075, 0.0375\}$.

The visual comparison shown in Figure 8 indicates an increase in σ_f values results in attribution maps with larger areas, while lower σ_f values result in a more concentrated explanation map. We report quantitative assessments in Figure 9. As the value of σ_f increases, there is an improvement in the metrics of AVG. % DROP and AVG. % INCREASE. However,

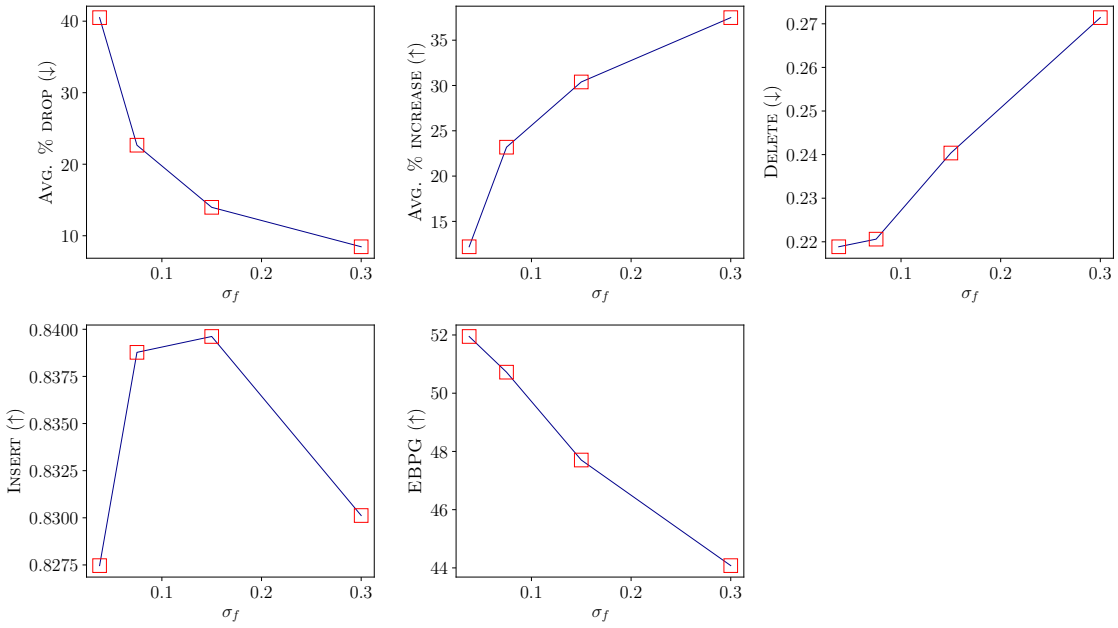

Figure 9: Quantitative assessment. Illustration of variation in performance of FovEx with respect to σ_f.

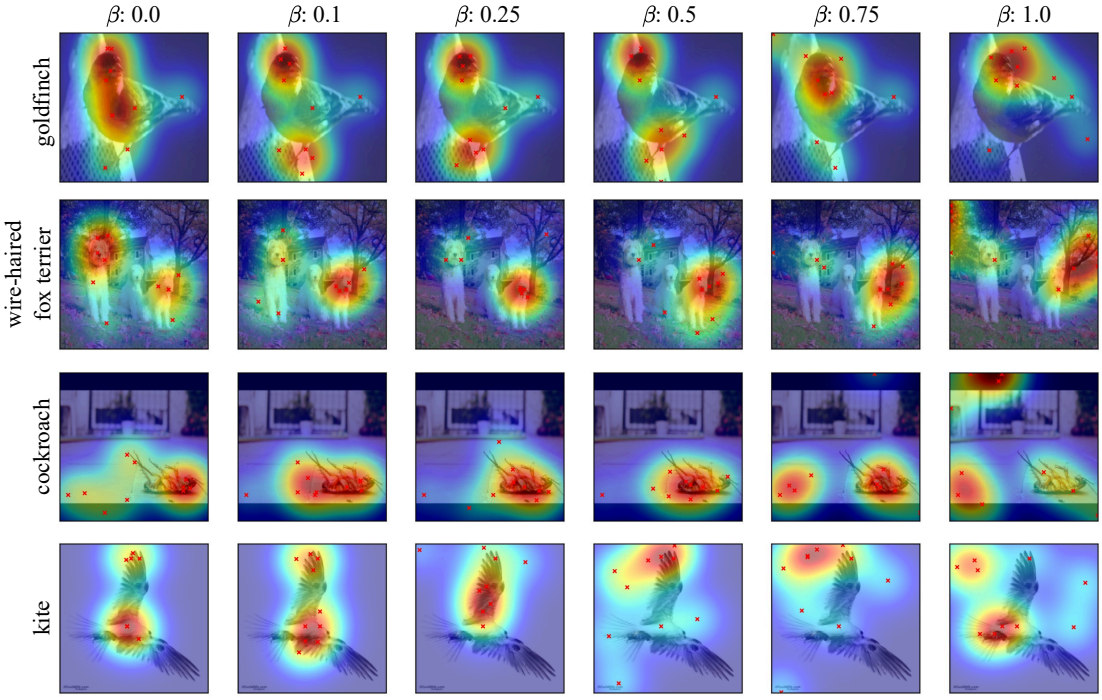

Figure 10: Qualitative assessment. Comparison of attribution maps generated using FovEx with different β values.

the metric of DELETE decreases. It is worth noting that an optimal σ_f value of 0.15 shows the maximum performance in the INSERT metric. Increasing σ_f leads to a wider spread of the attribution map, compromising localization.

Forgetting Factor (β). The forgetting factor $0 \leq \beta \geq 1$ regulates how much information is retained from previous fixations. In our study, we created attribution maps for various β values to investigate their effects. Specifically, we used the values $\{0.0, 0.1, 0.25, 0.5, 0.75, 1.0\}$. Visual comparison illustrated in Figure 10 reveals that the quality of

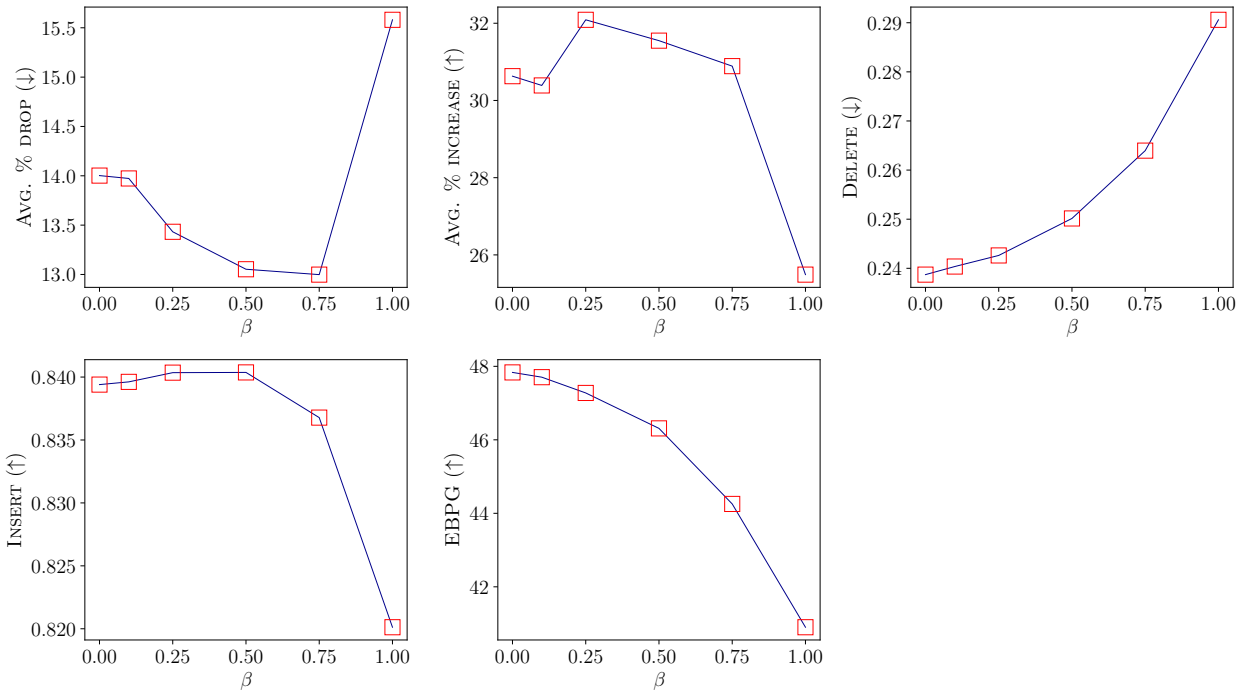

Figure 11: Quantitative assessment. Illustration of variation in performance of FovEx with respect to β.

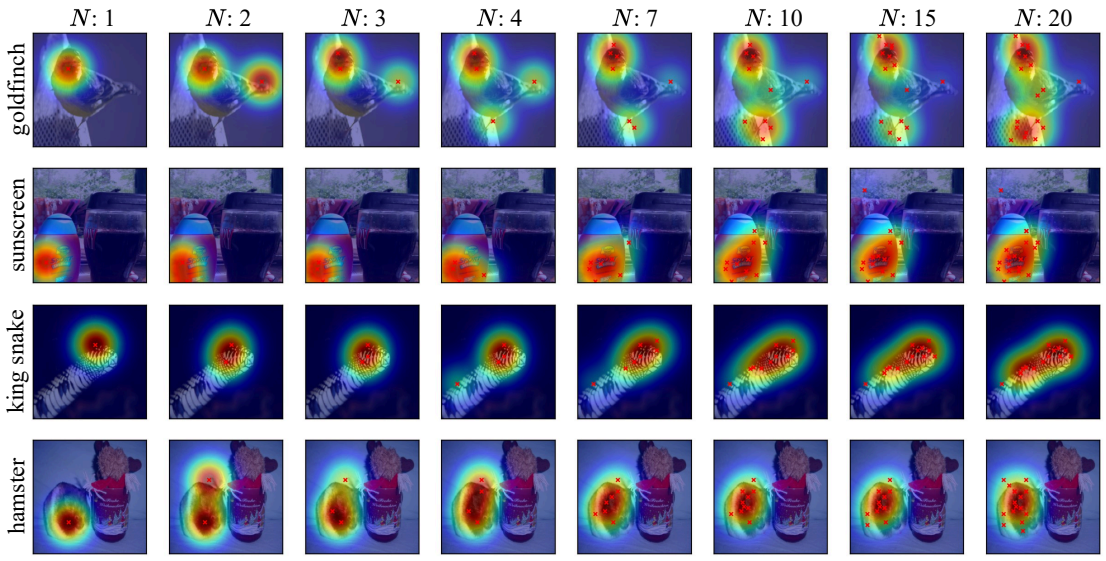

Figure 12: Qualitative assessment. Visual comparison of attribution maps generated using FovEx with different N values.

explanation maps decreases with an increase in the β value. However, the quantitative evaluation paints a nuanced picture as depicted in Figure 11.

Scanpath Length (N). N controls the length of the scanpath F, i.e., the number of fixation points computed to generate an attribution map E. To study the impact of N on the performance of FovEx, we generate attribution maps considering the following values for N, {1, 2, 3, 4, 7, 10, 15, 20}. As depicted in Figure 12 and Figure 13, both qualitative and quantitative (except EBPG and Time taken) performance improve with an increase in the value of N. However, the average time taken to produce an attribution map increases monotonically with N.

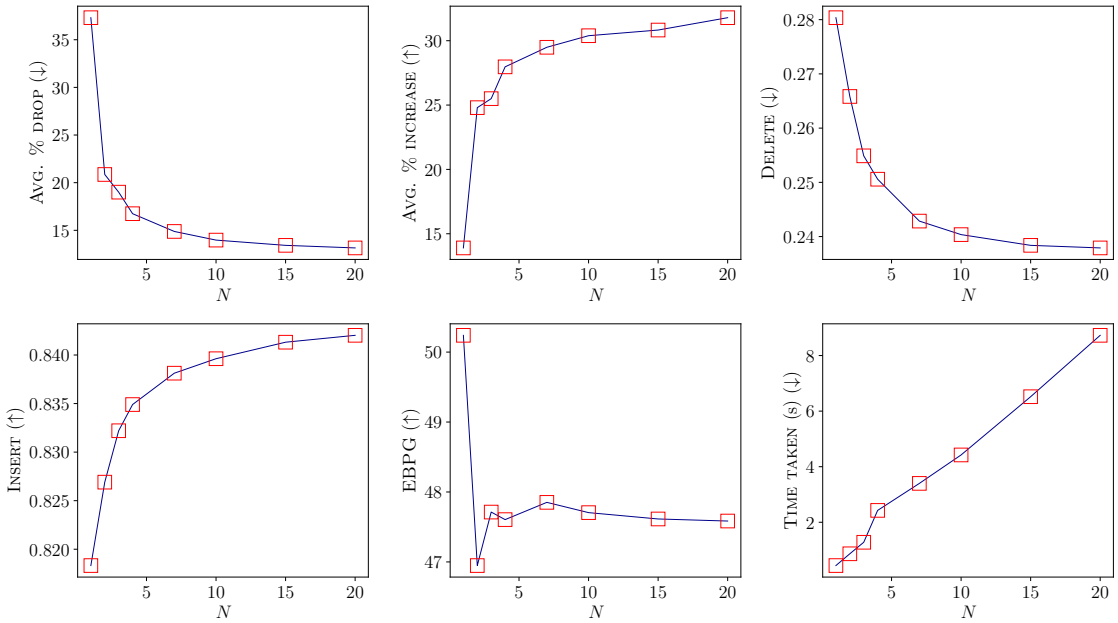

Figure 13: Quantitative assessment. Illustration of variation in performance of FovEx with respect to N.

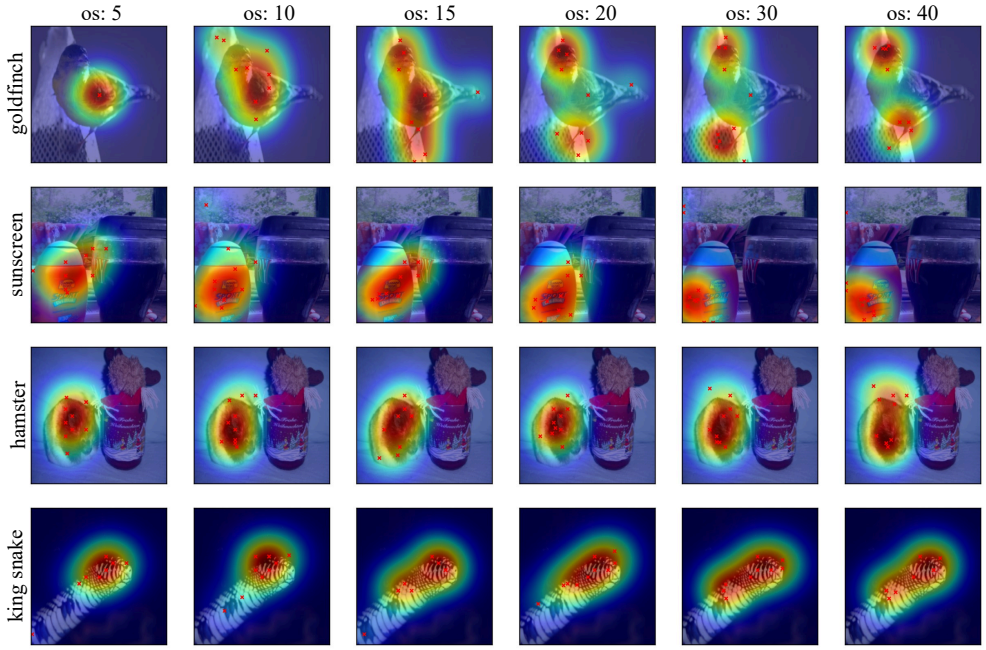

Figure 14: Qualitative assessment. Visual comparison of attribution maps generated using FovEx with different OS values.

Optimization Steps (OS). The OS parameter controls the number of iterations performed in the optimization scheme to find a fixation location. We generate attribution maps using the following values of OS, {5, 10, 15, 20, 30, 40}. The performance of FovEx improves with increment in OS values both qualitatively and quantitatively (except for DELETE and EBPG metric) as illustrated in Figure 14 and Figure 15. Similar to the effect of N with an increase in OS, the average time taken to produce an attribution map increases.

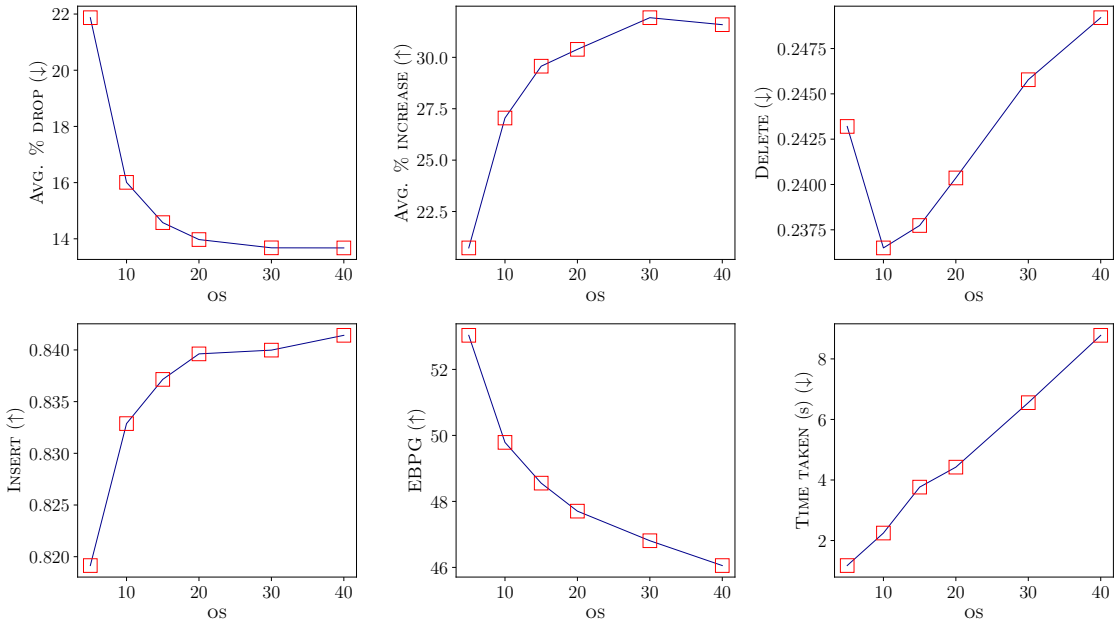

Figure 15: Quantitative assessment. Illustration of variation in performance of FovEx with respect to os.

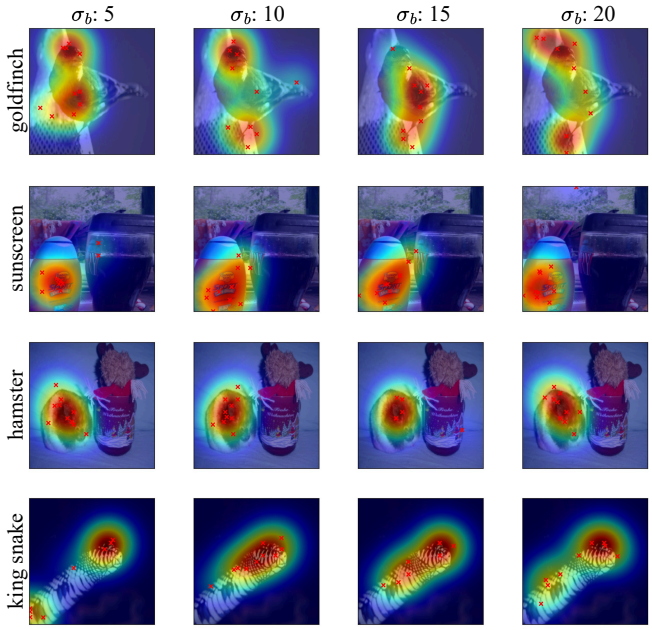

Figure 16: Qualitative assessment. Visual comparison of attribution maps generated using FovEx with different σ_b values.

Blur Sigma (σ_b). The σ_b parameter represents the standard deviation of the filter values used to create a blurred version of the input image which constitutes the peripheral region of the final foveated input image. We consider the following values for σ_b, {5, 10, 15, 20} for the experiments. Visual comparison in Figure 16 depicts plausible attribution maps generated at all σ_b values. However, the quantitative assessment illustrated in Figure 17 showcases an improvement in performance with increment in the value of σ_b except in the Insert metric.

Blur Filter Size (BFS). The BFS parameter represents the size of the Gaussian kernel used to create the blurred version of the input image. To study the effect BFS we consider the following values, {11, 21, 31, 41, 51}. Qualitative

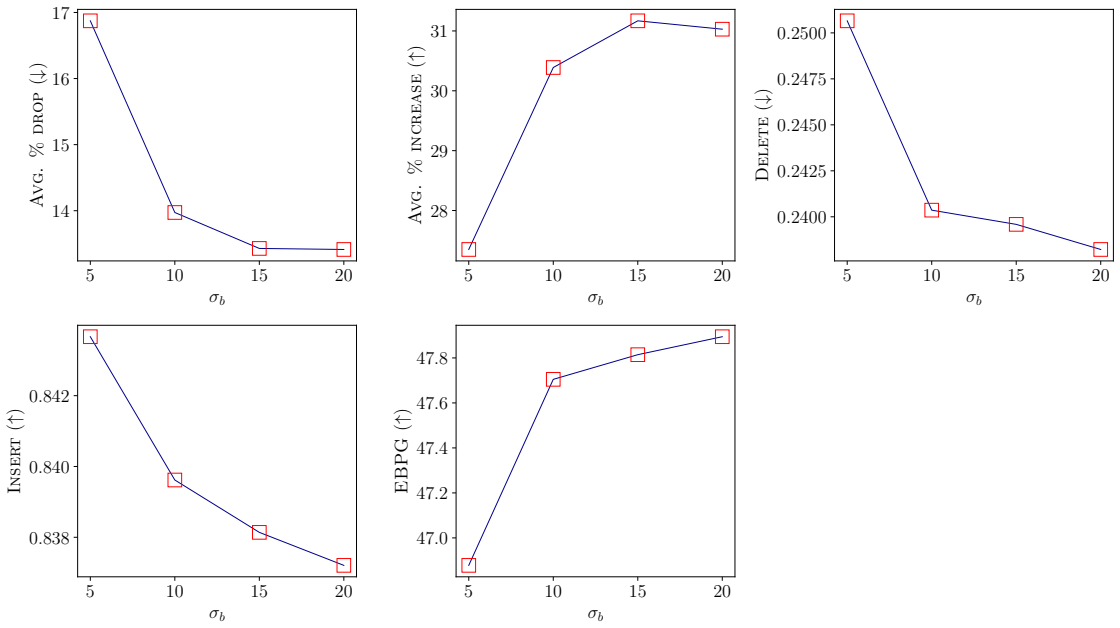

Figure 17: Quantitative assessment. Illustration of variation in performance of FovEx with respect to σ_b.

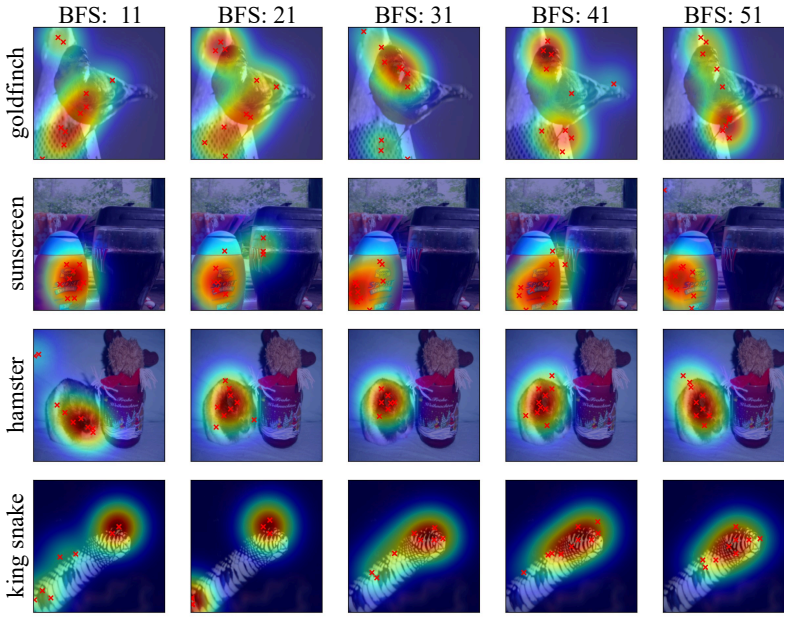

Figure 18: Qualitative assessment. Visual comparison of attribution maps generated using FovEx with different BFS values.

assessment illustrated in Figure 18, showcases visually slightly better attribution maps at BFS value of 41 than other values. Figure 19 illustrates the quantitative assessment. With an increase in BFS value the performance of FovEx increases except in INSERT metric.

5 Conclusion

We introduced FovEx, a novel explanation technique that integrates foveated human vision principles into the explanation generation process for black-box models. FovEx not only generates visually coherent explanation maps for both

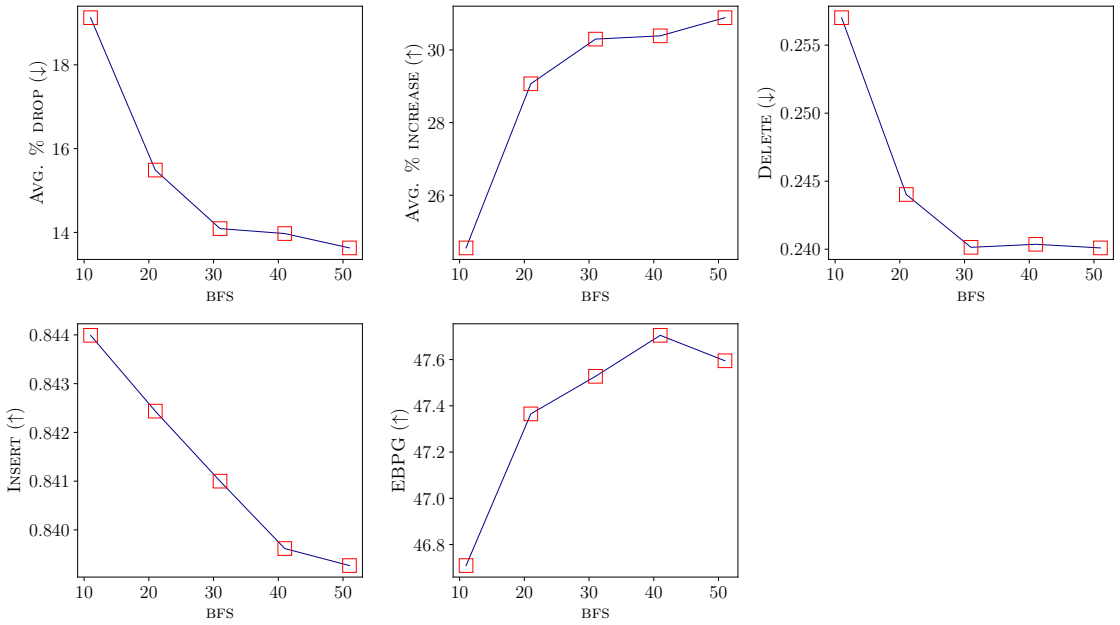

Figure 19: Quantitative assessment. Illustration of variation in performance of FovEx with respect to BFS.

transformer and convolution-based models but also possesses the ability to discern between different classes. It demonstrates superior performance in quantitative evaluations, outperforming competing methods for 4 out of 5 evaluation metrics for the ViT-B/16 model and for 3 out of 5 metrics for ResNet-50. Furthermore, incorporating biological foveated vision concepts allows FovEx to generate explanation maps that have a higher correlation to human gaze patterns on the considered dataset, as illustrated in Section 4.2, paving the way for a more accurate and intuitive understanding of model's decisions.

FovEx surpasses gradient-based and perturbation-based methods in all faithfulness metrics except the DELETE metric, for which it falls short in comparison to gradient-based methods. This is mostly due to FovEx's optimization scheme prioritizing the selection of image regions that are the most informative for the explanation, thereby offering a specific advantage in classification performance preservation, as reflected by other metrics such as INSERT. From an ethical perspective, we notice that the use of a small, potentially non-representative human gaze dataset might raise biases, and results on human alignment could not be representative of the diversity of a larger population. However, our method opens the door to better comparing model and human gaze pattern behavior.

Future work will extend the applicability of FovEx to models in various task domains, such as visual question answering and image captioning.

References

[1] Jiacheng Wang, Yueming Jin, and Liansheng Wang. Personalizing federated medical image segmentation via local calibration. In *European Conference on Computer Vision*, pages 456–472. Springer, 2022.

[2] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 762–780. Springer, 2020.

[3] Hironobu Fujiyoshi, Tsubasa Hirakawa, and Takayoshi Yamashita. Deep learning-based image recognition for autonomous driving. *IATSS research*, 43(4):244–252, 2019.

[4] Taotao Jing, Haifeng Xia, Renran Tian, Haoran Ding, Xiao Luo, Joshua Domeyer, Rini Sherony, and Zhengming Ding. Inaction: Interpretable action decision making for autonomous driving. In *European Conference on Computer Vision*, pages 370–387. Springer, 2022.

[5] Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R Qi, Xinchen Yan, Scott Ettinger, and Dragomir Anguelov. Motion inspired unsupervised perception and prediction in autonomous driving. In *European Conference on Computer Vision*, pages 424–443. Springer, 2022.

[6] Benjamin Maschler and Michael Weyrich. Deep transfer learning for industrial automation: A review and discussion of new techniques for data-driven machine learning. *IEEE Industrial Electronics Magazine*, 15(2): 65–75, 2021.

[7] Rahat Iqbal, Tomasz Maniak, Faiyaz Doctor, and Charalampos Karyotis. Fault detection and isolation in industrial processes using deep learning approaches. *IEEE Transactions on Industrial Informatics*, 15(5):3077–3084, 2019.

[8] Jacky Liang, Jeffrey Mahler, Michael Laskey, Pusong Li, and Ken Goldberg. Using dvrk teleoperation to facilitate deep learning of automation tasks for an industrial robot. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*, pages 1–8. IEEE, 2017.

[9] Virender Singh, Swati Singh, and Pooja Gupta. Real-time anomaly recognition through cctv using neural networks. *Procedia Computer Science*, 173:254–263, 2020.

[10] Jie Xu. A deep learning approach to building an intelligent video surveillance system. *Multimedia Tools and Applications*, 80(4):5495–5515, 2021.

[11] Monica Gruosso, Nicola Capece, and Ugo Erra. Human segmentation in surveillance video with deep learning. *Multimedia Tools and Applications*, 80:1175–1199, 2021.

[12] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

[13] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[14] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[15] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[16] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 397–406, 2021.

[17] Hoyoung Choi, Seungwan Jin, and Kyungsik Han. Adversarial normalization: I can visualize everything (ice). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124, June 2023.

[18] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020.

[19] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. Evaluating the impact of human explanation strategies on human-ai visual decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–37, 2023.

[20] Janet Hsiao and Antoni Chan. Towards the next generation explainable ai that promotes ai-human mutual understanding. In *XAI in Action: Past, Present, and Future Applications*, 2023.

[21] Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific reports*, 10(1):1411, 2020.

[22] Anna Volokitin, Gemma Roig, and Tomaso A Poggio. Do deep neural networks suffer from crowding? *Advances in neural information processing systems*, 30, 2017.

[23] Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

[24] Matteo Tiezzi, Simone Marullo, Alessandro Betti, Enrico Meloni, Lapo Faggi, Marco Gori, and Stefano Melacci. Foveated neural computation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 19–35. Springer, 2022.

[25] Aditya Jonnalagadda, William Yang Wang, BS Manjunath, and Miguel P Eckstein. Foveater: Foveated transformer for image classification. *arXiv preprint arXiv:2105.14173*, 2021.

[26] Leo Schwinn, Doina Precup, Björn Eskofier, and Dario Zanca. Behind the machine's gaze: Neural networks with biologically-inspired constraints exhibit human-like visual attention. *arXiv preprint arXiv:2204.09093*, 2022.

[27] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, 99: 101805, 2023.

[28] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9097–9107, 2019.

[29] Truong Thanh Hung Nguyen, Van Binh Truong, Vo Thanh Khang Nguyen, Quoc Hung Cao, and Quoc Khanh Nguyen. Towards trust of explainable ai in thyroid nodule diagnosis. *arXiv preprint arXiv:2303.04731*, 2023.

[30] Ian E. Nielsen, Dimah Dera, Ghulam Rasool, Ravi P. Ramachandran, and Nidhal Carla Bouaynaya. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, 2022.

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[32] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[33] 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):e0130140, 2015.

[34] Brian Kenji Iwana, Ryohei Kuroki, and Seiichi Uchida. Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation. In *International Conference on Computer Vision Workshops*, 2019.

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[36] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

[37] Jeong Ryong Lee, Sewon Kim, Inyong Park, Taejoon Eo, and Dosik Hwang. Relevance-cam: Your model already knows where to look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14944–14953, 2021.

[38] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[39] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[40] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[41] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2950–2958, 2019.

[42] Martina G. Vilas, Timothy Schaumlöffel, and Gemma Roig. Analyzing vision transformers for image classification in class embedding space, 2023.

[43] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[46] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[47] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

[48] Adriano Lucieri, Muhammad Naseer Bajwa, Andreas Dengel, and Sheraz Ahmed. Explaining ai-based decision support systems using concept localization maps. In *International Conference on Neural Information Processing*, pages 185–193. Springer, 2020.

[49] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

[50] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[51] Qiuxia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia*, 23:2086–2099, 2020.

[52] Ruoxi Qi, Yueyuan Zheng, Yi Yang, Caleb Chen Cao, and Janet H Hsiao. Explanation strategies for image classification in humans vs. current explainable ai. *arXiv preprint arXiv:2304.04448*, 2023.

[53] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

[54] Dario Zanca, Valeria Serchi, Pietro Piu, Francesca Rosini, and Alessandra Rufa. Fixatons: A collection of human fixations datasets and metrics for scanpath similarity. *arXiv preprint arXiv:1802.02534*, 2018.

[55] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[56] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[57] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406, October 2021.

Appendix A Performance on Various Models

We compare the performance of FovEx for three additional predictors ConvNeXt, ViT-B/16, and ViT-B/32 from torchvision. The ViT-B/16 model considered here is pre-trained on ImageNet-1K whereas the ViT-B/16 model in the main experiments is pre-trained on the ImageNet-21K dataset and fine-tuned on the ImageNet-1K dataset. We consider gradCAM [55], gradCAM++ [56], and Mean. Pert. [39] for comparison in the case of the ConvNeXt Base model. For ViT-B/16 and ViT-B/32 models we consider gradCAM and gradCAM++ for comparison. We employ randomCAM as a baseline method. We do not consider RISE [40] for ConvNeXt, ViT-B/16, and ViT-B/32 due to the high time requirements to generate the attribution maps. Similarly, due to high computational complexity, the Mean. Pert. method

Table 5: Quantitative assessment. Average metrics on the considered subset of ImageNet-1K validation dataset for ConvNeXt Base model. Bold terms denote the best performance and underlined terms represent the second best performance.

Eval. Name	FovEx	grad CAM	grad CAM++	Mean. Pert.	random CAM
AVG. % DROP (↓)	**59.6592**	80.9472	<u>78.2356</u>	87.0630	86.6874
AVG. % INCREASE (↑)	**9.7699**	3.4699	<u>3.6999</u>	1.6999	0.7499
DELETE (↓)	0.2142	0.2098	<u>0.1907</u>	**0.0739**	0.2638
INSERT (↑)	**0.1833**	0.1548	<u>0.1606</u>	0.1422	0.1478
EBPG (↑)	48.1485	**53.1640**	<u>49.4505</u>	43.3129	30.9518

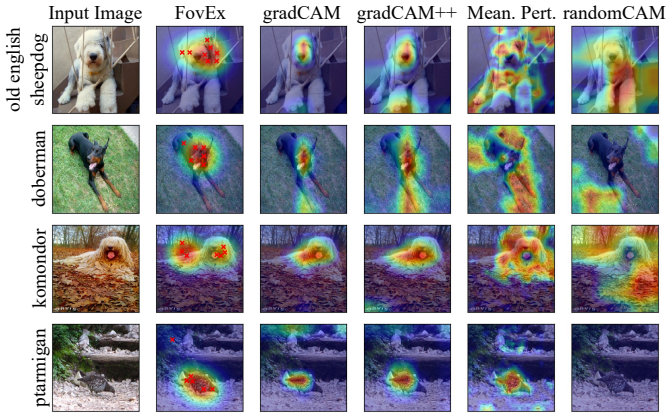

Figure 20: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, and Mean. Pert. for ConvNeXt Base model.

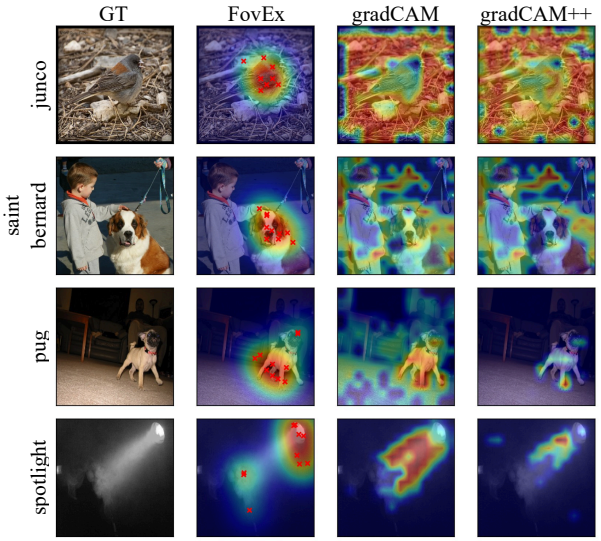

Figure 21: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, and randomCAM for ViT-B/16 model.

is not considered for ViT-B/16 and ViT-B/32. Transformer-specific methods like GAE [57] and Cls. Emb. [42] methods are not utilized due to the need to change the pre-existing implementation from torchvision.

We report qualitative assessments in Figure 20, Figure 21, and Figure 22 for ConvNeXt Base, ViT-B/16 and ViT-B/32 models respectively. For all three cases, FovEx creates noise-free and consistent attribution maps. Table 5, Table 6, and Table 7 depict the qualitative evaluation results for the ConvNeXt Base model, ViT-B/16 and ViT-B/32 model respectively. FovEx maintains state-of-the-art performance across all the models by outperforming all other XAI methods in 3 out of 5 metrics in the case of ConvNeXt and in 5 out of 5 metrics in the case of ViT-B/16 and ViT-B/32 models.

Appendix B Default Parameter Values.

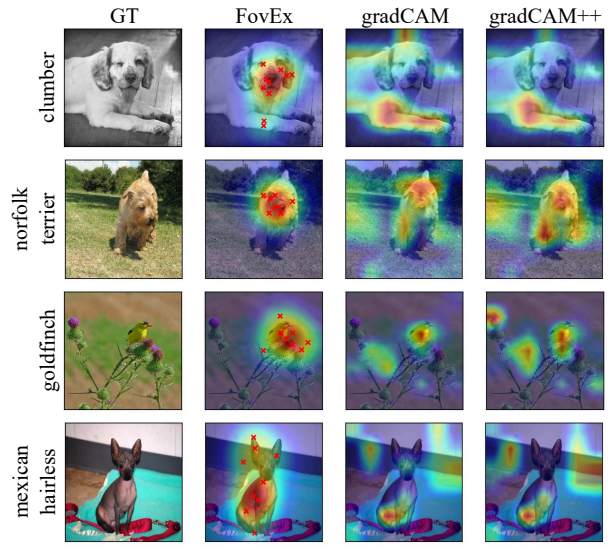

Figure 22: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, and randomCAM for ViT-B/32 model.

Table 6: Quantitative assessment. Average metrics on the considered subset of ImageNet-1K validation dataset for ViT-B/16 model. Bold terms denote the best performance and underlined terms represent the second best performance.

Eval. Name	FovEx	grad CAM	grad CAM++	random CAM
AVG. % DROP (↓)	**48.2258**	66.9577	75.6945	91.4881
AVG. % INCREASE (↑)	**17.1490**	4.9299	3.5299	0.6490
DELETE (↓)	**0.0886**	0.0913	0.1101	0.1283
INSERT (↑)	**0.3647**	0.2956	0.2964	0.2777
EBPG (↑)	**48.1485**	39.6095	38.0989	36.0899

Table 7: Quantitative assessment. Average metrics on the considered subset of ImageNet-1K validation dataset for ViT-B/32 model. Bold terms denote the best performance and underlined terms represent the second best performance.

Eval. Name	FovEx	grad CAM	grad CAM++	random CAM
AVG. % DROP (↓)	**44.0460**	64.1220	70.1500	85.8250
AVG. % INCREASE (↑)	**32.6890**	16.8690	13.7690	5.3499
DELETE (↓)	**0.1034**	0.1261	0.1241	0.1565
INSERT (↑)	**0.4992**	0.4293	0.4303	0.3773
EBPG (↑)	**48.4390**	40.1740	40.9060	37.6540

Table 8: Parameter Values. These values are used in Section 4.1 and Appendix A.

Parameter Name	Value
RR	True
σ_f	0.15
β	0.1
N	10
OS	20
σ_b	10
BFS	41
λ	0.1

Appendix C Additional Visualizations of Explanation Maps

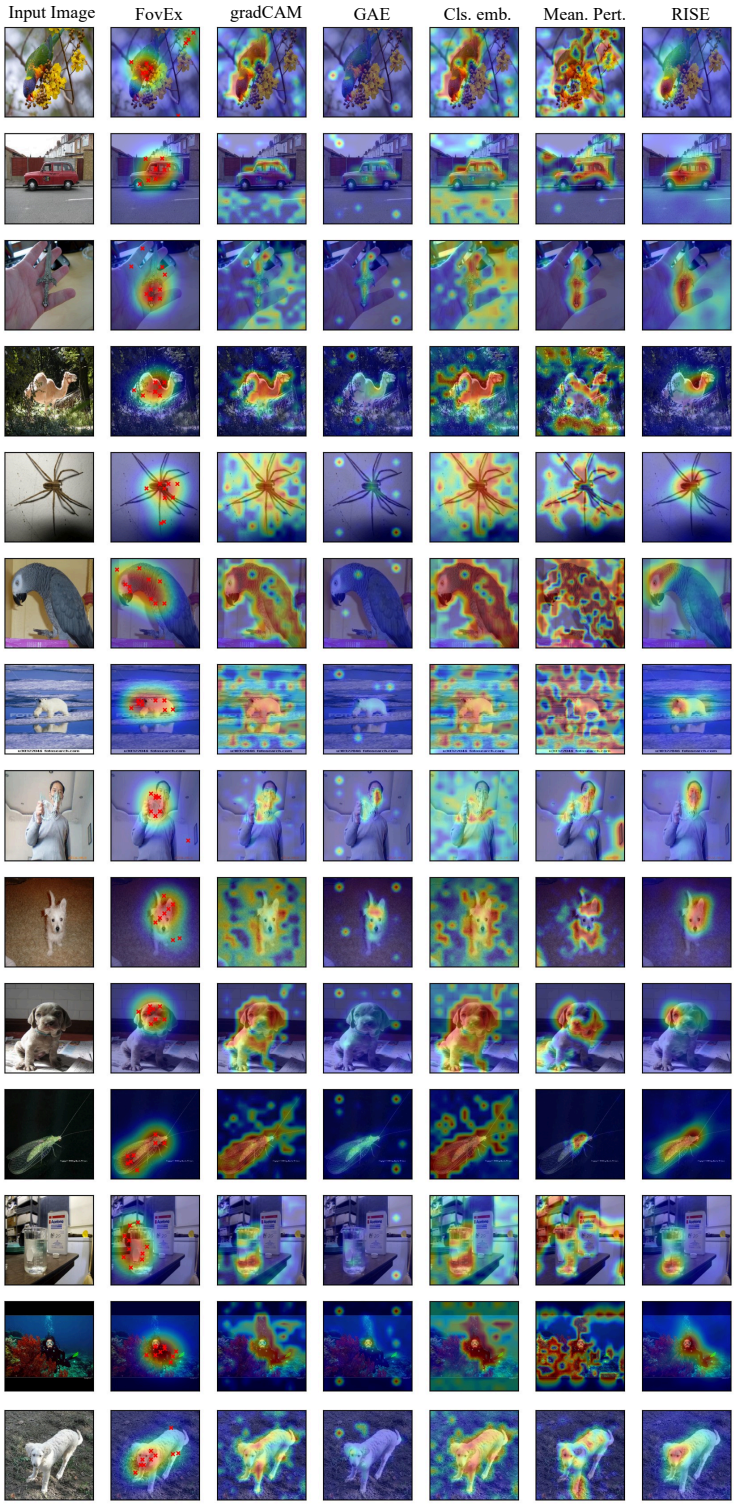

Figure 23: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, GAE, Cls. Emb., Mean. Pert., and RISE for ViT-B/16 model used in Section 4.1.

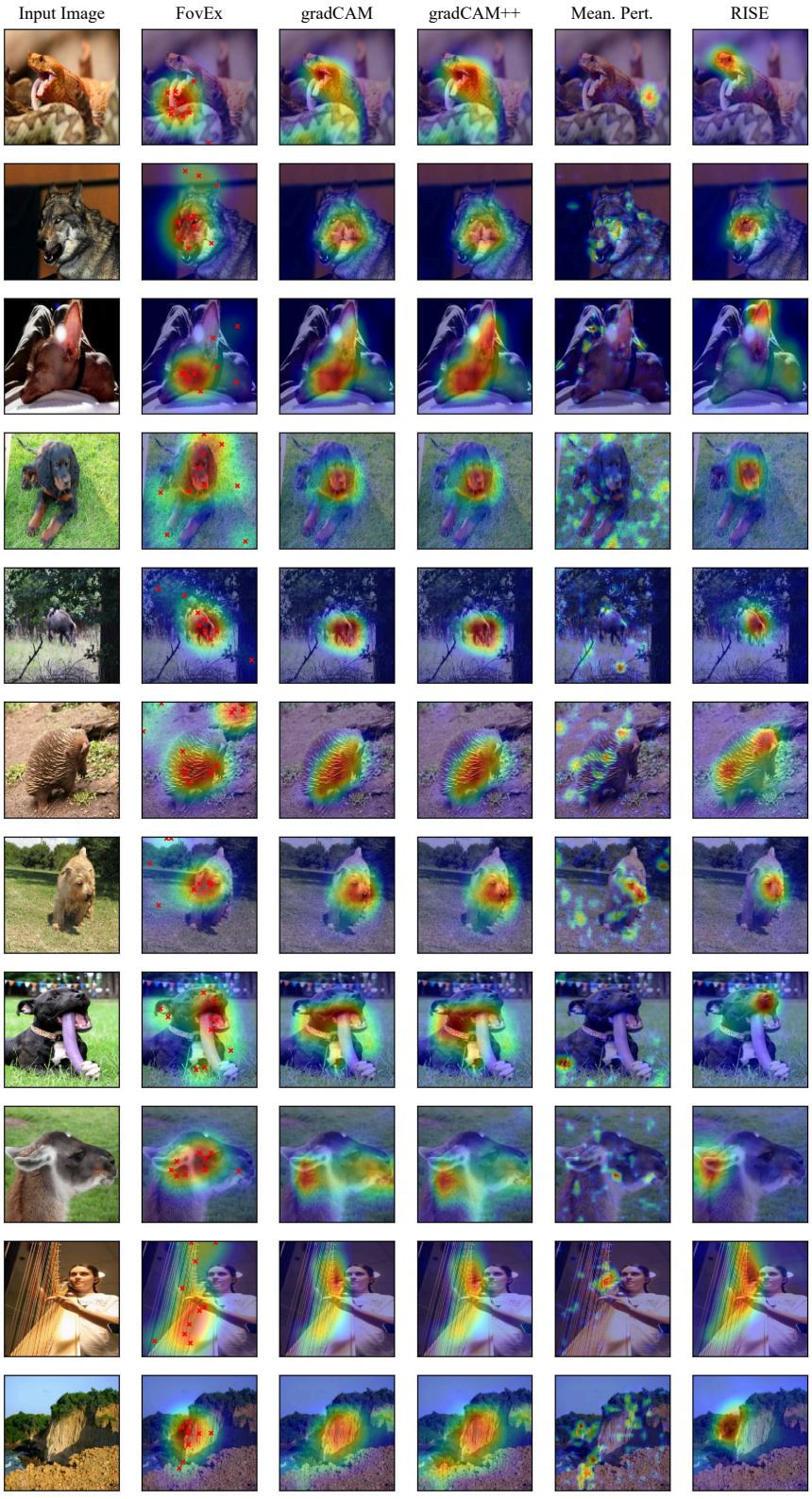

Figure 24: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, Mean. Pert., and RISE for ResNet-50 model used in Section 4.1.

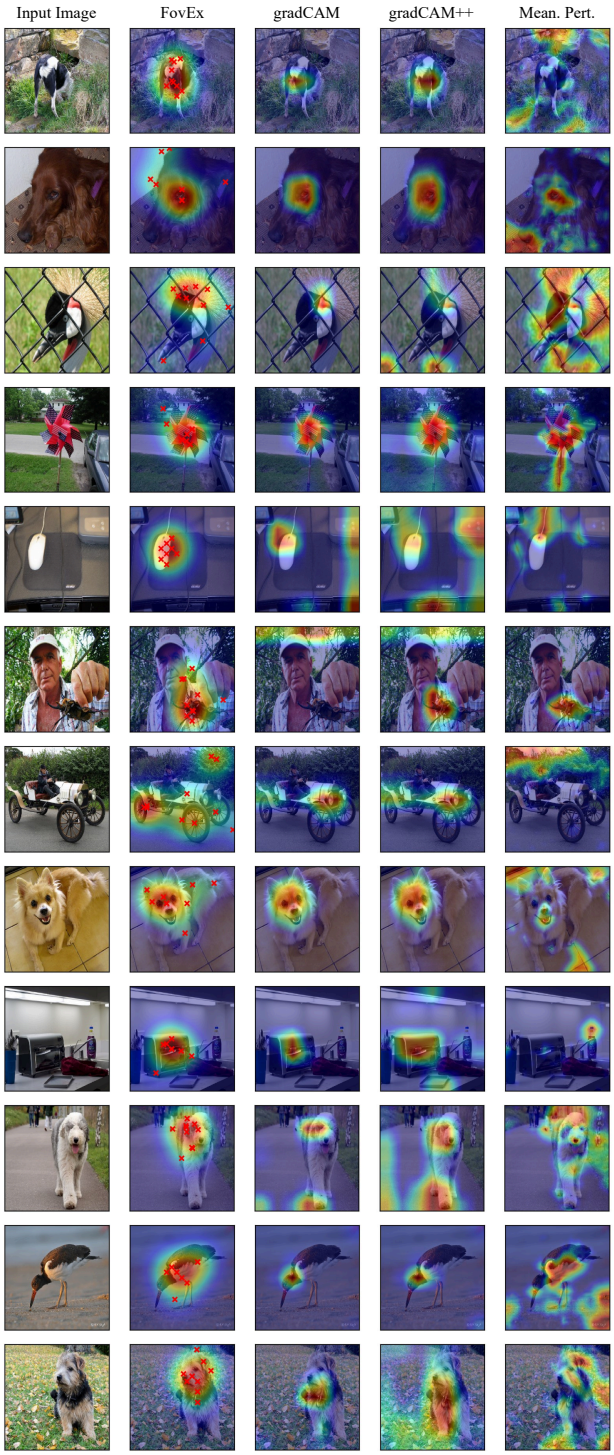

Figure 25: Qualitative assessment. Visual comparison of attribution maps generated using FovEx, gradCAM, gradCAM++, and Mean. Pert. for ConvNeXt model used in Appendix A.