Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

P-TAME: Explain Any Image Classifier with Trained Perturbations

Mariano V. Ntrougkas[1,2], Vasileios Mezaris[1], Senior Member, IEEE, and Ioannis Patras[2], Senior Member, IEEE

[1]Information Technologies Institute / CERTH, Thessaloniki, Greece
[2]Queen Mary University of London, London, UK

Corresponding author: Mariano V. Ntrougkas (email: ntrougkas@iti.gr).

This work was supported by the EU Horizon Europe program under grant agreement 101070190 AI4TRUST.

ABSTRACT The adoption of Deep Neural Networks (DNNs) in critical fields where predictions need to be accompanied by justifications is hindered by their inherent black-box nature. This paper introduces P-TAME (Perturbation-based Trainable Attention Mechanism for Explanations), a model-agnostic method for explaining DNN-based image classifiers. P-TAME employs an auxiliary image classifier to extract features from the input image, bypassing the need to tailor the explanation method to the internal architecture of the backbone classifier being explained. Unlike traditional perturbation-based methods, which have high computational requirements, P-TAME offers an efficient alternative by generating high-resolution explanations in a single forward pass during inference. We apply P-TAME to explain the decisions of VGG-16, ResNet-50, and ViT-B-16, three distinct and widely used image classifiers. Quantitative and qualitative results show that P-TAME matches or outperforms previous explainability methods, including model-specific ones. Code and trained models are available at https://github.com/IDT-ITI/P-TAME.

INDEX TERMS Explainable AI, Deep Learning, Computer Vision, Attention Mechanism

I. INTRODUCTION

ADVANCES in deep neural networks (DNNs) over the past decade have been tremendous. However, a persistent challenge is the lack of DNN explainability [1]. DNNs are often referred to as "black-box" models because they do not provide users with insights into their decision-making process, posing a significant barrier to their wider adoption in important application domains such as healthcare, journalism, and law enforcement, where the ability to justify decisions is a critical requirement [2], [3]. Consequently, there is a growing interest in developing methods to make DNN decisions more understandable to users, i.e., in developing eXplainable Artificial Intelligence (XAI) methods [1].

Within this research domain, a dominant direction to advancing the explainability of DNN image classifiers is to generate saliency maps [4], which highlight the regions of the input image that are most relevant to the decision of the DNN. Saliency maps (a.k.a. explanation maps; Fig. 1) can help users understand why a DNN made a particular decision and can also be used to identify potential biases in the decision-making process [5].

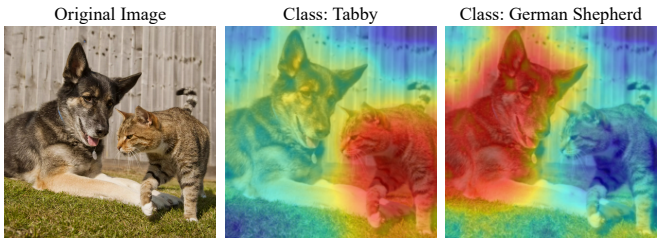

FIGURE 1. Example class-specific explanations produced by the P-TAME method. The image classifier whose predictions are being explained is the ViT-B-16 model, using the default weights from torchvision.

Several classes of methods have been proposed to generate saliency maps for DNNs, including gradient-based [6], [7], perturbation-based [8]–[10], and response-based [11]–[15] methods. Gradient-based methods compute the gradient of the output with respect to the input image and use it to generate the saliency map. They suffer from the vanishing gradient problem and can be noisy and unreliable [16]. Additionally, the most widely used methods in this category, Grad-CAM and Grad-CAM++ [6], [7] require the extraction

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

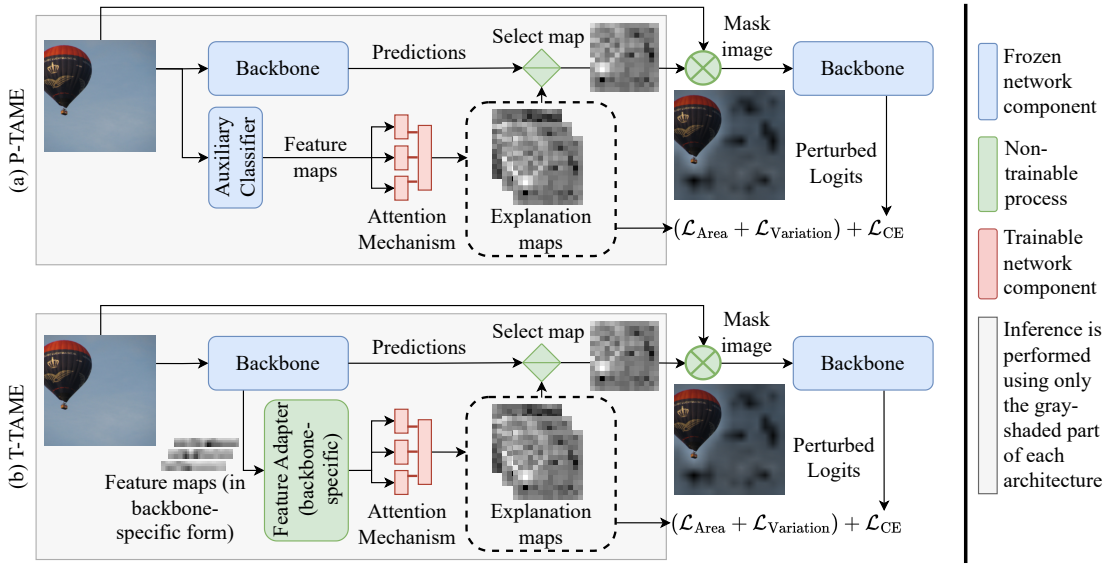

FIGURE 2. **Overview of the proposed P-TAME method (a), displaying the pipeline for both the training and inference stages. In this illustration, the main difference between the P-TAME method (a) and T-TAME (b) is evident: In P-TAME, no intermediate feature maps are extracted from the backbone (i.e., the DNN classifier whose decisions we aim to explain). Instead, an auxiliary classifier is employed to extract feature maps from the input image.**

of intermediate feature maps from the network being explained, thus needing to be adapted to the DNN architecture of interest. Perturbation-based methods generate saliency maps by perturbing the input image and observing the change in the output. They are more robust and reliable than gradient-based methods but are computationally expensive at the inference stage, which is a critical constraint in e.g., real-time edge applications in industrial inspection, where explanations must be generated in a resource-constrained environment [17]. Furthermore, they are not always model-agnostic; e.g., the widely used Score-CAM [8] relies on extracting intermediate feature maps (similarly to Grad-CAM, Grad-CAM++). Response-based methods, e.g., CAM [13], generate saliency maps by combining the intermediate feature maps of the DNN to predict the saliency map. They are, by definition, model-specific and often also make use of perturbations, making them computationally expensive.

To address the limitations of previous methods, i.e., the noisy explanations produced by gradient-based methods and the computationally expensive process of using perturbations, TAME [14] and T-TAME [15] (Transformer-compatible Trainable Attention Mechanism for Explanations) proposed a new paradigm for explaining the decisions of DNNs by generating saliency maps with an attention mechanism. The attention mechanism learns to combine the intermediate feature maps from multiple layers of the DNN to predict the saliency map. The quality of the produced explanation maps is generally on par with perturbation-based methods, while avoiding the need for multiple forward passes during inference. A limitation of many previous methods that persists, however, is that feature maps need to be extracted from the DNN that is being explained. Additionally, depending on the DNN architecture, these feature maps may need to be adapted to the T-TAME attention mechanism. Thus, TAME and T-TAME are not model-agnostic, in contrast to many perturbation-based approaches.

In this work, we propose P-TAME (Perturbation-based Trainable Attention Mechanism for Explanations), an attention-based XAI method that generates saliency maps directly from the input images using an auxiliary classifier, without the need to extract and process intermediate feature maps from the DNN being explained. Since P-TAME is model-agnostic, it can be applied to any DNN image classifier. It produces saliency maps by learning to perturb the input image to highlight the regions most relevant to the decision of the DNN being explained; and, after training, produces explanations in a single forward step. The performance of P-TAME is evaluated, both quantitatively [7], [18] and qualitatively, on three popular image classifiers: VGG-16 [19], ResNet-50 [20], and ViT-B-16 [21] trained on ImageNet [22]. Experimental comparisons demonstrate that P-TAME rivals state-of-the-art (SoA) perturbation methods in explanation quality, without needing multiple forward passes during inference. We provide P-TAME as an open-source library to support adoption and further XAI research.

In summary, the contributions of this work are as follows:

- We propose P-TAME, a method that employs an auxiliary classifier to extract feature maps, which are then processed by an attention mechanism to generate explanation maps. P-TAME is model-agnostic, thus can be easily applied to any DNN image classifier.
- We evaluate P-TAME quantitatively on three popular image classifiers with very different architectures trained on the ImageNet dataset, and we compare it against T-TAME and other SoA methods.

II. RELATED WORK

Humans have long explained and justified their actions, a core aspect of how they relate, cooperate, and build trust [23]. Conversely, the inability of current AI systems to provide justifications for their actions hampers trust in their decisions. There are many ways to increase trust and transparency of AI systems, but in this work we will focus on techniques that produce explanations for the decisions of image classifiers. The form of these explanations varies, depending on the nature of the data which the AI system is designed to work on. For image classifiers, the most common form for explanations is a feature attribution map, also known as explanation map (Fig. 1). These explanations are local, as opposed to global, because they explain a single decision of the classifier, instead of describing how an image classifier reaches its decisions in general. In addition to explanation maps, recently, methods that aim to explain a classifier's decisions using a set of semantic concepts have been introduced, e.g., [24]–[26]. However, due to the considerably different form of the produced explanations, the latter methods are not comparable to XAI methods producing explanation maps.

This section establishes a brief taxonomy of XAI methods (for a more comprehensive taxonomy, refer to [27]) and describes notable XAI methods for image classifiers. To produce explanations for an image classifier, we can employ an intrinsically explainable AI model (e.g. [28]), called an ante-hoc explainable model, or apply an XAI method to a trained model without modifying it. The latter methods are called post-hoc and have the advantage of being applicable to SoA image classifiers without trading off performance for explainability. Post-hoc methods are further divided into model-agnostic and model-specific methods. Model-agnostic methods only require access to the model input and output, while model-specific methods require access to the model architecture and may have specific requirements to be applicable. A relative drop in the number of new post-hoc explainability methods for image classifiers was observed after the release of the Vision Transformer (ViT) [21], due to a partial shift in focus from developing model-agnostic methods, to exploring the self-attention maps of Vision Transformers for explainability [29], [30]. However, many newer Transformer-based image classifiers, such as [31], [32], make ViT-specific explainability methods inapplicable, leaving only a few model-agnostic approaches like RISE [9] to be applicable to these and any other classifier.

We can further categorize post-hoc XAI methods for image classifiers by their approach to producing explanation maps. Gradient-based methods, like Grad-CAM and Grad-CAM++ [6], [7], produce explanations using the model's gradients. Grad-CAM [6] produces explanations using a weighted sum of the feature maps of the final layer before classification. The weights are computed via global average pooling of the gradient for each feature map with respect to the output class. These methods are simple and intuitive but face gradient-related issues, including noise and saturation from the activation functions [16], [33]. Additionally, because they utilize intermediate feature maps, they are not model-agnostic. Perturbation-based approaches observe how the model's outputs vary when the input is distorted. These methods can be model-agnostic, like RISE [9], LIME [34] and SHAP [35], or utilize intermediate feature maps, such as Score-CAM [8], Opti-CAM [36], TAME [14], and T-TAME [15]. RISE [9] generates random masks and uses them to mask the input image. The output confidence scores are used as weights in the weighted sum of the masks. LIME [34] trains a surrogate model based on the model's responses after perturbations of the input, but relies on superpixels for image explanations, being too computationally expensive to produce less coarse explanations. SHAP [35] approximates Shapley values with SHAP scores, which represent the contribution of each input feature to the final decision of the classifier. Computing the exact Shapley values for complex neural network-based models is not feasible for high-dimensional data (e.g., images), and the approximated SHAP scores produce misleading information about relative feature importance [37], resulting in poor performance in XAI metrics measuring faithfulness. Score-CAM [8] uses the DNN's final layer's feature maps, claiming that they represent better perturbations. Opti-CAM [36], like Score-CAM, uses the final layer's feature maps, but trains a weight vector during inference to maximize the model's confidence. CAM [13] is a purely response-based method, using only the final layer's feature maps and the global average pooling layer's output to produce explanation maps, which constrains its application to very specific architectures. SISE [11] and Ada-SISE [12] blur the boundaries between the categories of gradient-, perturbation-, and response-based methods by using the gradients of the model's predictions, combining intermediate feature maps, and using them to perturb the input. TAME [14] and T-TAME [15] (the latter being applicable not only to convolutional neural networks (CNNs), as TAME is, but also to Vision Transformer-based architectures) probe the model during training, utilizing feature maps from multiple layers and learning weights to combine them. During inference, they produce explanations without perturbations, lowering computational requirements. Therefore, they are trainable response-based approaches; however, their attention mechanism is trained prior to inference (in contrast to Opti-CAM).

The proposed P-TAME is a trainable perturbation-based approach, and unlike T-TAME (and TAME), it is also model-agnostic, imposing no constraints on the backbone architecture. Additionally, while P-TAME needs to be trained separately for each backbone to be explained, training is performed with a standardised procedure that is identical regardless of the backbone. P-TAME is performant during inference, requiring only a single forward pass to produce explanations, and is easily trainable and applicable to any DNN-based image classifier. We note that the explanation

maps produced by P-TAME are not related to the attention maps produced by Visual Transformers [29], as P-TAME is specifically trained to produce explanation maps that highlight regions which are important for the backbone's decision, while ViT attention maps are a byproduct of the multi-head attention, whose value as explanations is unclear [38].

III. P-TAME

A. METHOD OVERVIEW

The process of yielding explanations for the predictions of image classifiers with P-TAME involves two main steps. The first step is to train an attention mechanism that generates explanation maps from feature maps. In contrast to T-TAME, feature maps are never directly extracted from the backbone network (the DNN whose decisions should be explained); instead, they are produced by an auxiliary classifier (whose weights are also frozen). Thus, the P-TAME method is model-agnostic: only the input images and the backbone's output predictions are required. The second step involves using the trained attention mechanism to directly produce class-specific explanations for the backbone's predictions.

The pipeline of the proposed framework is illustrated side-by-side with the T-TAME pipeline in Fig. 2, highlighting the main difference between the two methods, which is the introduction of the auxiliary classifier in P-TAME.

B. DEFINITIONS

Consider an image classifier network (a.k.a. backbone) $f\colon \mathcal{X} \to \mathbb{R}^C$ that maps an input image $x \in \mathcal{X}$ to a vector of logits $y = (y)_x = f(x) \in \mathbb{R}^C$, where $\mathcal{X}$ is the space of images and C is the number of classes. We denote the c-th element of y as y_c. Let $c^* = \arg\max y$ be the model-truth class, i.e., the model's prediction, which can be contrasted with a ground-truth class provided by a labeled dataset. Additionally, consider an auxiliary image classifier network $f_{\text{aux}}\colon \mathcal{X} \to \mathbb{R}^C$. The auxiliary classifier f_{aux} is constrained to only CNN-based architectures, because they produce three-dimensional feature maps. We denote the feature map extracted from layer l of the auxiliary classifier as $F_l \in \mathbb{R}^{d_l \times w_l \times h_l}$. Here, d_l, w_l, and h_l are the number of channels, height, and width of the feature map, respectively. The attention mechanism of P-TAME takes as input multiple feature maps from different layers of the auxiliary classifier, to improve the resolution of the produced explanation maps, based on the findings of [15]. Let $\mathcal{A}(F_L) = E$ be the attention mechanism, where F_L is the set of feature maps extracted from $L = \{l_1, l_2, \ldots, l_s\}$ layers, and $E \in [0,1]^{C \times w_E \times h_E}$ the class-specific explanation maps. Finally, we denote by $R = w_E \cdot h_E$ the resolution of the explanation maps.

C. AUXILIARY CLASSIFIER & ATTENTION MECHANISM

The auxiliary classifier, a CNN pretrained on the same dataset as the backbone (e.g. ResNet-18 [20], see Section IV-A for experimentation details), extracts features that follow a predictable pattern: deeper layers capture semantically rich features, while earlier layers detect simple patterns or edges [39]. These features are three-dimensional, spatially consistent with the input image, and straightforward to process. The P-TAME attention mechanism combines feature maps from various layers of the auxiliary classifier, which differ in channel count and spatial resolution. Using these feature maps, it generates explanations that highlight the most salient input regions according to the backbone. This adaptation involves processing each feature map individually and combining them to produce class-specific explanation maps, as illustrated in Fig. 3a. We note that the auxiliary classifier's role is to produce well-structured and diverse input-specific feature maps for the attention mechanism to combine, not to mirror the backbone's logic. The trained P-TAME attention mechanism is responsible for reflecting the backbone's reasoning.

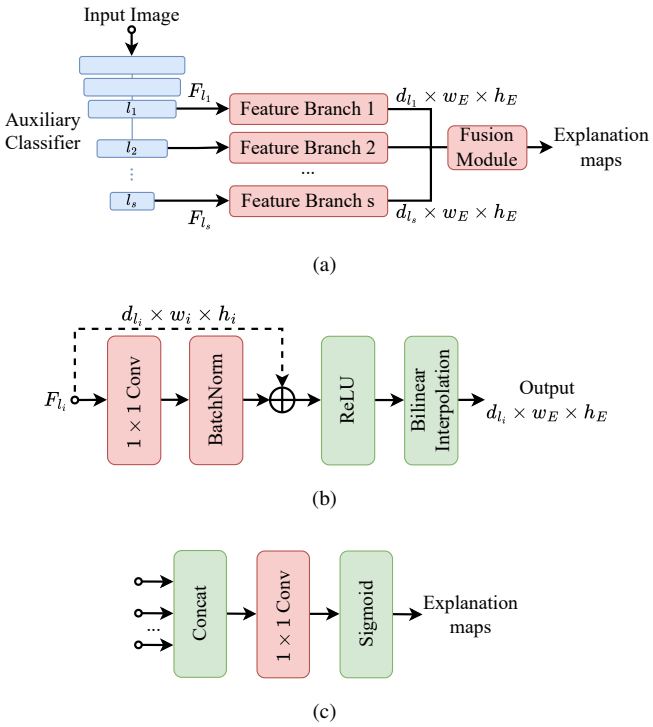

FIGURE 3. Structure of the attention mechanism of P-TAME (also used in T-TAME, though with different input). (a) Overview of the auxiliary classifier and the attention mechanism, (b) detailed structure of a feature branch of the attention mechanism, (c) detailed structure of the fusion module of the attention mechanism. The same color coding as in Fig. 2 is used to denote frozen / non-trainable / trainable components.

Feature maps F_L from different layers of the auxiliary classifier are processed individually through a feature branch comprising a 1×1 convolution layer, batch normalization, a skip connection, an activation function, and bilinear interpolation (Fig. 3b). Bilinear interpolation upscales smaller feature maps to match the resolution of the largest feature map. While feature maps extracted from deeper

layers typically have lower resolutions, some architectures produce feature maps of equal resolution (e.g., architectures using inverted residual blocks), making bilinear interpolation necessary only when resolutions differ. All feature maps are scaled to the largest resolution, matching the resolution of the final explanation maps (R). The processed feature maps are concatenated and passed through a 1×1 convolution layer and a sigmoid activation function, producing class-specific explanation maps (Fig. 3c). The sigmoid activation ensures that the resulting explanation maps have values in $[0, 1]$.

D. TRAINING REGIME

The attention mechanism we defined has to be trained to correctly combine the input feature maps into meaningful class-specific explanation maps. The auxiliary classifier's weights are frozen, so only the attention mechanism's weights need to be trained. This is done in a self-supervised manner, similarly to T-TAME (Fig. 2). Specifically, images from the dataset used to train the backbone f are input to both f and the components of P-TAME: the auxiliary classifier f_{aux} and the attention mechanism $\mathcal{A}$. During training, to measure how salient the explanations produced by P-TAME for the training image x are, we first select the explanation E_{c^*} (corresponding to the model truth class c^*) and use it to mask the image:

$$x_m = x \odot \text{up}_{\text{bilinear}}(E_{c^*}), \tag{1}$$

where $\text{up}_{\text{bilinear}}(\cdot)$ refers to bilinear interpolation, and $\odot$ refers to the Hadamard product. Then, the masked image is input a second time to the backbone to produce new predictions $f(x_m) = (y)_{x_m}$. The masking procedure removes features that should be of low relevance to the classifier's prediction. After removing these features, we expect the confidence in the prediction to rise, as this is the basic premise of visual attention [40]. We measure the fidelity of the explanations through the response of the model with the cross-entropy loss:

$$\mathcal{L}_{\text{CE}}(c^*, (y)_{x_m}) = -\log((y_{c^*})_{x_m}), \tag{2}$$

For this, we use the model-truth class c^* instead of the classifier's original prediction $(y_{c^*})_x$, due to the training instability of soft cross-entropy [29]. The use of model-truth resembles many knowledge distillation methods (e.g., [29], [41]), and we can view the training of P-TAME as a form of distillation of the backbone model's reasoning.

With a naive minimization of the above loss, the all-ones mask $x_m = x \odot \mathbf{1} = x$ would be the trivial solution. To avoid this, we add a second loss term $\mathcal{L}_{\text{Area}}$, which penalizes the produced explanations based on how activated they are. For calculating this loss term, we consider not only the explanation for class c^* but also for other classes, specifically for a uniformly sampled subset S of $\{0, ..., C-1\}$ with $c^* \in S$ and $|S| = \lambda_{\text{rand}}$, where λ_{rand} is a hyperparameter. We use this subset S instead of all C classes to avoid excessive calculations. Thus, we define $\mathcal{L}_{\text{Area}} = \frac{1}{|E_S|}\sum E_S^{\lambda_{\text{area}}}$, where λ_{area} is a hyperparameter and $|E_S| = |S| \cdot R$ is the total number of elements in the explanation maps E_S.

Additionally, we want to encourage simpler explanations. To minimize explanation complexity, we penalize the spatial variation within each explanation map belonging to the same subset S:

$$\mathcal{L}_{\text{Variation}} = \frac{1}{|E_S|}\sum_{c\in S}\left(\|\nabla_j E_{c,j,k}\|^2 + \|\nabla_k E_{c,j,k}\|^2\right), \tag{3}$$

where $\nabla_j E_{c,j,k} = E_{c,j+1,k} - E_{c,j,k}$ and $\nabla_k E_{c,j,k} = E_{c,j,k+1} - E_{c,j,k}$ represent the spatial derivatives of E_c. Finally, the loss function used to train P-TAME is:

$$\begin{aligned}\mathcal{L}(c^*, (y)_{x_m}, E_S) =& \lambda_1 \mathcal{L}_{\text{CE}}(c^*, (y)_{x_m}) \\ &+ \lambda_2 \mathcal{L}_{\text{Area}}(E_S) \\ &+ \lambda_3 \mathcal{L}_{\text{Variation}}(E_S),\end{aligned} \tag{4}$$

where $\lambda_{\{1,2,3\}}$ are hyperparameters. Here we can observe that P-TAME is wholly agnostic to the specific architecture of the image classifier f that is being explained.

E. INFERENCE

During inference, only one forward pass is required to compute explanation maps, as illustrated in Fig. 2. The image is input to the backbone classifier to generate a prediction and to the auxiliary classifier to extract feature maps. Then, the feature maps are processed by the trained attention mechanism to generate class-specific explanation maps.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

We perform a comprehensive evaluation of P-TAME by comparing it both quantitatively and qualitatively against SoA explainability methods across 3 backbone image classifiers: VGG-16 [19], ResNet-50 [20], and ViT-B-16 [21]. For measuring explanation quality, we adopt evaluation measures that are widely used in the domain. We also report the resolution of the produced explanation maps before rescaling and measure the computational requirements of different explainability methods by reporting the number of forward passes required to produce an explanation. Additionally, we perform a sanity check, per [16], to determine that the explanations produced by P-TAME are sensitive to the parameters of the backbone model. Furthermore, we perform an ablation study examining the effects of different choices of auxiliary classifiers both quantitatively and qualitatively. In the latter ablation, we compare three lightweight image classifiers: ResNet-18 [20], MobileNetV3 [43], and MnasNet [44]. Besides assessing differences in the explanation quality, we also compare the computation requirements imposed by each auxiliary classifier (measured in GFLOPs) and quantify how the features extracted from each different layer of the auxiliary classifiers contribute to the final explanation maps.

Dataset: We use the ImageNet ILSVRC 2012 dataset [45]. The training subset of it (1,281,167 images) is used for training P-TAME, while two subsets of 2000 images each from

TABLE 1. **Comparison of P-TAME with SoA methods using the AD, IC, MoRF and LeRF measures.**

Backbone	Method	AD↓			IC↑			ROAD (AUC)			
		100%	50%	15%	100%	50%	15%	MoRF↓	LeRF↑	R↑	Fwd Passes↓
VGG-16 (acc@1: 71.59% [42])	Grad-CAM [6]	32.12%	58.65%	84.15%	22.10%	9.50%	2.20%	21.34%	65.76%	49	**1**
	Grad-CAM++ [7]	30.75%	54.11%	82.72%	22.05%	11.15%	3.15%	22.57%	64.54%	49	**1**
	RISE [9]	8.74%	42.42%	78.70%	51.30%	17.55%	4.45%	22.72%	**69.25%**	49	4000
	Score-CAM [8]	27.75%	45.60%	75.70%	22.80%	14.10%	4.30%	22.12%	66.66%	49	512
	Ablation-CAM [10]	34.87%	49.23%	76.96%	19.25%	11.45%	3.65%	20.69%	66.95%	49	2048
	Opti-CAM [36]	**2.23%**	42.66%	87.97%	**85.91%**	20.78%	2.18%	26.24%	61.21%	49	50
	T-TAME [15]	9.33%	36.50%	**73.29%**	50.00%	**22.45%**	**5.60%**	**18.55%**	66.93%	784	**1**
	P-TAME	7.11%	**33.39%**	76.06%	49.06%	22.17%	4.76%	24.78%	68.34%	**3136**	**1**
ResNet-50 (acc@1: 76.13% [42])	Grad-CAM [6]	13.61%	29.28%	78.61%	38.10%	23.05%	3.40%	24.80%	73.38%	49	**1**
	Grad-CAM++ [7]	13.63%	30.37%	79.58%	37.95%	23.45%	3.40%	25.95%	72.34%	49	**1**
	RISE [9]	11.12%	36.31%	82.05%	46.15%	21.55%	3.20%	**23.42%**	**73.74%**	49	8000
	Score-CAM [8]	11.01%	**26.80%**	78.72%	39.55%	24.75%	3.60%	27.01%	72.10%	49	512
	Ablation-CAM [10]	13.58%	30.33%	79.62%	37.05%	22.30%	3.50%	25.78%	72.23%	49	8192
	Opti-CAM [36]	**1.27%**	38.49%	90.00%	**90.87%**	24.60%	1.79%	32.83%	62.97%	49	50
	T-TAME [15]	7.81%	27.88%	78.58%	54.00%	**27.50%**	**4.90%**	24.61%	68.89%	784	**1**
	P-TAME	8.35%	28.95%	**77.53%**	50.00%	24.85%	4.81%	26.13%	71.27%	**3136**	**1**
ViT-B-16 (acc@1: 81.07% [42])	Grad-CAM [6]	37.19%	40.74%	73.11%	12.75%	12.30%	5.40%	27.65%	71.92%	196	**1**
	Grad-CAM++ [7]	57.21%	72.77%	92.51%	5.55%	4.85%	0.80%	46.98%	64.35%	196	**1**
	RISE [9]	38.09%	44.20%	77.50%	15.35%	14.50%	4.85%	36.85%	**76.28%**	49	8000
	Score-CAM [8]	35.50%	42.16%	80.86%	8.90%	10.55%	2.95%	32.25%	62.65%	196	768
	Ablation-CAM [10]	38.09%	44.20%	77.50%	15.35%	14.50%	4.85%	33.30%	72.27%	196	768
	Opti-CAM [36]	**0.15%**	67.29%	93.36%	**98.07%**	13.29%	1.88%	47.62%	54.51%	196	50
	T-TAME [15]	8.19%	23.64%	72.89%	38.35%	40.40%	9.40%	**24.66%**	74.97%	196	**1**
	P-TAME	7.50%	**19.63%**	**62.69%**	47.47%	**43.45%**	**11.86%**	33.89%	73.01%	**3136**	**1**

the ILSVRC 2012 evaluation set are used as our validation and testing sets, respectively. The number of image classes (in this dataset and the pre-trained backbones and auxiliary classifiers used in our experiments) is $C = 1000$.

Models: For the backbone image classifiers VGG-16 [19], ResNet-50 [20] and ViT-B-16 [21], we use their ImageNet-pretrained instances available in the `torchvision` library [42]. These classifiers represent three very distinct evolutionary phases in the field of DNN-based image classification, each introducing significant architectural shifts w.r.t. their predecessors, i.e., the 2-dimensional convolution layer, the skip connection, and the multi-head attention layer. A ResNet-18 [20] model, also pretrained on ImageNet and retrieved from `torchvision`, is used as our auxiliary classifier, chosen because it strikes a good balance between performance and computational requirements. Feature maps are extracted from the outputs of the last four residual blocks of ResNet-18. Other choices of auxiliary classifiers are considered in the ablation study (Section IV-C).

Training: We train P-TAME's attention mechanism on the ImageNet dataset for one epoch, using a batch size of 64 images, the largest batch size that our GPU can support (following [46]). In our experiments, we observed that further training did not improve the performance of the method. We use the AdamW optimizer [47] and the OneCycleLR learning rate scheduler [48], setting the maximum learning rate to either 10^{-4} or 10^{-3}. The hyperparameter λ_{rand} is set equal to the batch size. Prior to this training, to determine appropriate values for the hyperparameters introduced in the loss function (Eq. 4), we utilize Bayesian optimization, specifically the BoTorch framework [49]. Bayesian optimization is a well-established technique for serial optimization of costly-to-evaluate black-box functions, such as the training and evaluation of a neural network. Bayesian optimization involves conducting several trials, and for each one, we train for a single epoch and then evaluate it using the MoRF and LeRF measures (see "Evaluation measures", below) computed on the validation set. The search space is greatly compacted by constraining the loss term weights by $\sum_i \lambda_i = 1$, allowing λ_1 and λ_2 to take values in the range $[0, 1]$ with the condition that $\lambda_1 + \lambda_2 < 1$ and setting $\lambda_3 = 1 - \lambda_1 - \lambda_2$. Also, λ_{area} is allowed to take a value from set $\{0.5, 1, 2\}$. With only 5 initial random trials and 15 subsequent trials of Bayesian optimization (i.e., 20 trials in total), the hyperparameter optimization procedure converges. The exact parameters to reproduce the reported results are included in the released source code.

Evaluation measures: For evaluating explainability methods for image classifiers, the most crucial aspect of explanations that we want to quantify is their "faithfulness", or how much they align with the image classifier that is being explained. The approach most frequently used in the domain is to perturb the input image, using the explanation map as a mask of the image to observe how the confidence in the original prediction changes. We use 8 measures to capture "faithfulness". The most widely employed measures, Average Drop (AD) and Increase in Confidence (IC), are defined as [7]:

$$\text{AD}(v) = \sum_x \frac{\max\{0, (y_{c^*})_x - (y_{c^*})_{x_{m(v)}}\}}{\Upsilon}, \quad (5)$$

$$\text{IC}(v) = \sum_x \frac{\text{int}\left((y_{c^*})_{x_{m(v)}} > (y_{c^*})_x\right)}{\Upsilon}, \quad (6)$$

where Υ represents the number of test images. Here, $x_{m(v)}$ is the masked image, with a threshold applied to the mask

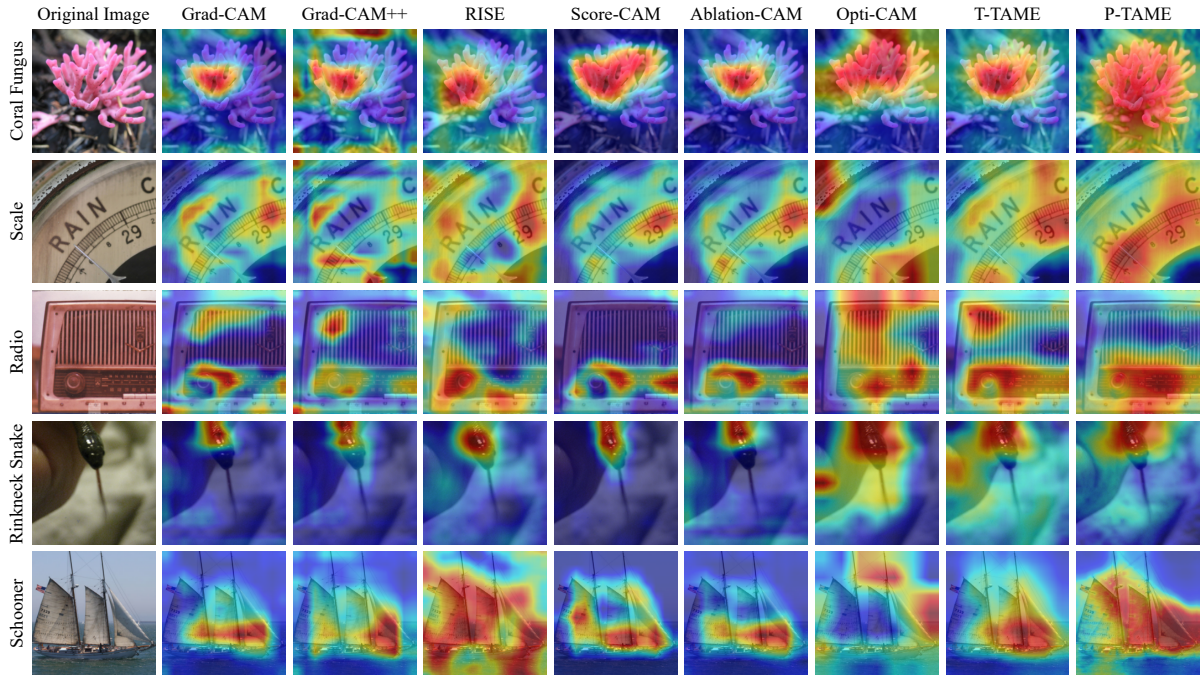

FIGURE 4. **Explanation maps produced by different methods for the ResNet-50 [20] backbone. The model truth class is shown on the left.**

to select the top $v\%$ highest-valued pixels of the explanation map E_{c^*}. We also use the MoRF and LeRF measures [18]:

$$\text{MoRF}(v) = \sum_x \frac{\mathbb{I}((c^*)_{x_{\hat{m}(v)}} = (c^*)_x)}{\Upsilon}, \tag{7}$$

$$\text{LeRF}(v) = \sum_x \frac{\mathbb{I}((c^*)_{x_{\check{m}(v)}} = (c^*)_x)}{\Upsilon}, \tag{8}$$

where $\mathbb{I}()$ is an indicator function that returns 1 if the condition is true and 0 otherwise, $x_{\hat{m}(v)}$ and $x_{\check{m}(v)}$ denote the image masked with a binary mask which selects the top $v\%$ highest or lowest valued pixels of the explanation map, respectively. The masking procedure is a type of image infilling, described in [18]. We threshold the mask at percentages $(10\%, 20\%, 30\%, 40\%, 50\%, 70\%, 90\%)$ as in [18] to assess the effectiveness of the explanation in ranking pixel importance. The area under the curve of the resulting accuracies is computed to aggregate the results from the various thresholds. A low MoRF indicates that the explanation map correctly identifies the most significant image regions for the prediction, while a high LeRF signifies accurate identification of the least significant regions. MoRF and LeRF are independent of mask distribution and rely solely on pixel ranking, with the infilling procedure mitigating input distribution shifts, which particularly impact CNNs [50].

B. QUANTITATIVE RESULTS AND COMPARISONS

In Table 1, our proposed P-TAME method is compared with the following SoA methods: Grad-CAM [6], Grad-CAM++ [7], RISE [9], Score-CAM [8], Ablation-CAM [10], and T-TAME [15]. We selected these specific methods because they

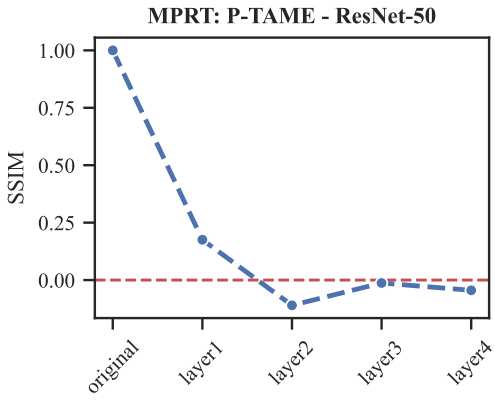

FIGURE 5. **Model Parameter Randomization Test (MPRT) for P-TAME on the ResNet-50 [20] backbone. We observe a sharp drop in SSIM as the backbone's layers (indicated on the horizontal axis) are randomized, showcasing that P-TAME passes the sanity check for saliency maps [16].**

are among the most widely used and performant methods of their respective class (gradient-, perturbation-, and response-based approaches). From the results, we observe that for the ViT-B-16 backbone, we obtain top performance in the AD and IC measures, except for the $v = 100\%$ threshold, which is dominated by Opti-CAM across different backbones. However, Opti-CAM exhibits the worst performance in the more challenging AD(15%), IC(15%), and ROAD measures. For the CNN models VGG-16 and ResNet-50, we obtain near-top scores for the AD and IC measures, competing in performance only with T-TAME and the model-agnostic perturbation method RISE. In the MoRF and LeRF measures, which signal if the ordering of pixels by importance is

TABLE 2. Ablation study: different choices of auxiliary classifier.

Auxiliary Classifier	ResNet-18 [20]	MobileNet-V3 [43]	MnasNet [44]	ResNet-50 [20]	ResNet-18* [20]
AD 100%↓	8.35%	11.15%	14.11%	**7.81%**	9.13%
IC 100%↑	50.00%	42.26%	38.29%	**54.00%**	44.59%
AD 50%↓	28.95%	36.59%	43.24%	**27.88%**	36.55%
IC 50%↑	24.85%	19.64%	16.22%	**27.50%**	18.40%
AD 15%↓	**77.53%**	79.81%	83.83%	78.58%	84.27%
IC 15%↑	4.81%	4.56%	2.68%	**4.90%**	2.78%
MoRF↓	26.13%	**24.28%**	26.77%	24.61%	28.32%
LeRF↑	**71.27%**	69.13%	66.76%	68.89%	64.53%
R↑	**3136**	49	49	784	**3136**
GFLOPs↓	46.42	**24.97**	26.89	104.89	46.42
Layer 1	6.73%	11.11%	3.67%	-	6.72%
Layer 2	13.44%	11.10%	32.09%	14.31%	13.48%
Layer 3	26.68%	11.06%	31.55%	28.61%	26.81%
Layer 4	53.15%	66.72%	32.69%	57.08%	52.99%

correct, P-TAME provides mixed results. This is mostly caused by the fact that the explanation maps produced by P-TAME have a much higher resolution, and providing a good ordering of R pixels is much simpler for lower resolutions. This is further elucidated in Section IV-D. Still, the fact that P-TAME generates explanation maps in a single forward step and can be applied to any image classifier architecture is a significant advantage compared to more computationally intense methods (e.g., RISE) or more restrictive feature map extraction methods (e.g., Grad-CAM, T-TAME).

In Fig. 5, the outcome of a Model Parameter Randomization Test (MPRT), a sanity check to determine whether an XAI method is sensitive to the backbone's parameters [16], is shown for P-TAME applied to the ResNet-50 backbone. Structure Similarity Index Measure (SSIM) values are calculated between the explanations produced for the pretrained ResNet-50 model and the explanations produced after randomizing all its parameters up to the layer named in the x-axis, following a bottom-up approach as recommended in [51]. An SSIM value near zero denotes no similarity between the compared explanation maps. We observe a sharp decrease in SSIM immediately, demonstrating that P-TAME is highly sensitive to the randomization of the backbone and thus passes the sanity check for saliency maps.

C. ABLATIONS

In Table 2, we examine different auxiliary classifiers for explaining the ResNet-50 backbone, initially comparing our choice of ResNet-18 with MobileNetV3 [43] and MnasNet [44] (again, models pretrained on ImageNet, retrieved from [42]). We observe that smaller auxiliary classifiers offer computational advantages but produce coarser explanation maps due to lower feature resolution. This also results in modest improvements in MoRF and LeRF for MobileNetV3, as it is easier to produce explanation maps with $49 = 7^2$ elements than with $3136 = 56^2$ elements. Overall, however, using ResNet-18 outperforms using any of the other two models, indicating a clear tradeoff between compute and explanation quality in selecting the auxiliary classifier. We also note that the contribution of feature maps extracted from different layers to the final explanation maps varies greatly across classifiers. These contributions, calculated by processing the fusion module's trained weights (Fig. 3c) and grouping them based on which feature branch they correspond to (Fig. 3a), show that the deeper layer's feature maps consistently contribute more. In ResNet-18, contributions increase steadily with deeper layers, while MobileNetV3 and MnasNet show near-equal contributions across layers. This difference is due to the architectures of MobileNetV3 and MnasNet, which use strided convolutions followed by inverted residual blocks, in contrast to the typical residual blocks found in ResNet-18. Inverted residual blocks are computationally efficient but yield feature maps with fewer channels and small spatial dimensions, making it harder for P-TAME to transform these feature maps into class-specific explanation maps.

In the case that the auxiliary classifier is identical to the backbone being explained (i.e., in this case, ResNet-50), P-TAME degenerates to T-TAME (Fig. 2; in this experiment, feature maps from 3 layers are used, following the T-TAME protocol [15]). Using the much larger ResNet-50 model as an auxiliary classifier yields moderate improvements across most measures, but more than doubles the computational requirements (GFLOPs). Note that T-TAME is model-specific; i.e., when the backbone also serves as the feature map generator (instead of using an auxiliary classifier), changes in the backbone being explained necessitate architectural changes in the attention mechanism of the explanation method.

In a final ablation, we experiment with using as auxiliary classifier a ResNet-18 model trained on Imagenette [52], a small subset of ImageNet containing just 10 classes and 1% of the original images. This ablation examines the extent to which it is necessary to use an auxiliary classifier pretrained on the same dataset as the backbone being explained. The Imagenette-pretrained auxiliary classifier is denoted as "ResNet-18*" in Table 2. We observe from these results near-top performance (compared to using other auxiliary classifiers) for the AD and IC at higher thresholds (Table 1), but a significant drop at the 15% threshold and for the ROAD measures. This demonstrates that despite using an auxiliary classifier trained on a much smaller dataset, P-TAME continues to produce meaningful explanations; however, achieving SoA results requires a feature map generator trained on the backbone's original training data.

D. QUALITATIVE RESULTS

In Fig. 4, explanation maps produced for the ResNet-50 backbone using P-TAME and the SoA methods of Table 1 are shown, following the findings of [53] on the importance of complementing quantitative evaluation with qualitative analysis. We select the ResNet-50 backbone for this qualitative comparison because it is one of the most widely used CNN architectures, and most of the compared explainability methods were developed for CNNs. We observe that P-

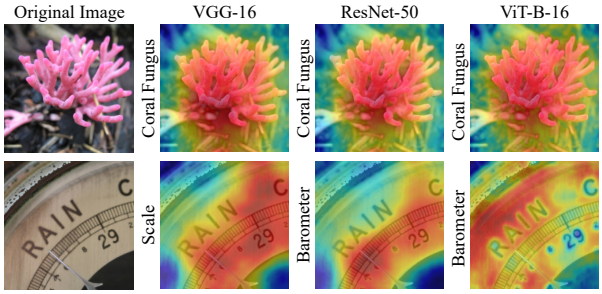

FIGURE 6. Explanation maps produced for the VGG-16 [19], ResNet-50 [20], and ViT-B-16 [21] backbones. The model-truth class of the original image according to each backbone is shown on the left.

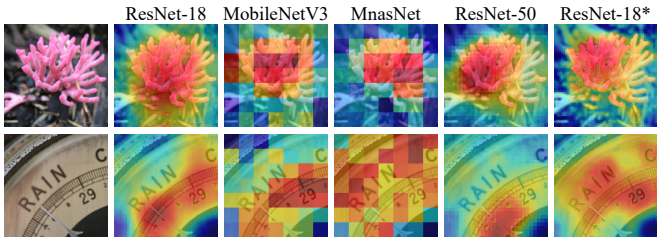

FIGURE 7. Explanation maps produced for the ResNet-50 [20] backbone, using different auxiliary classifiers. The model-truth class of these images is "coral fungus" and "barometer", respectively. To illustrate the raw resolution of the generated explanations, we upscale the explanation maps using nearest-neighbor interpolation (avoiding the smoothing effects of bilinear interpolation, used in other figures).

TAME produces the most activated explanation maps, followed by T-TAME and RISE. P-TAME correctly highlights the entire class when it can be localized (rows 1, 4, 5). In cases where the class cannot be localized, P-TAME correctly highlights salient features, in line with methods that directly make use of features extracted from the backbone. Along with the good quantitative results in Table 1, this shows that P-TAME produces high-quality explanation maps in a single forward pass without requiring any backbone architecture-specific tailoring to extract and process feature maps. The only other model-agnostic method, RISE, besides requiring 8000 forward passes to produce the shown explanation maps, produces noisier results, especially in cases where the class is not easily localizable (rows 2, 3, 5).

In Fig. 6, we compare explanation maps produced for our three backbones (VGG-16, ResNet-50, and ViT-B-16) using P-TAME. For the first image, which shows a localizable class, the explanation maps are similar across backbones. However, for the second image, whose class cannot be easily localized to a specific region of the image, the ViT-B-16 backbone, the most performant model out of the three in terms of classification performance (see 1st column of Table 1), shows the highest level of detail in its explanation. For example, the number "29" in the second image is shown to have low importance for the model-truth prediction. For less performant models, like VGG-16, the explanations show much less detail, even though the resolution of the explanation map is the same as for ViT-B-16. This indicates a performance-explainability trade-off, i.e., that a higher-performing classifier can support the generation of more detailed explanations for it.

In Fig. 7, explanation maps produced for the ResNet-50 backbone using different auxiliary classifiers are shown. To demonstrate the differences in the resolution of the explanation maps, we use nearest-neighbor upscaling instead of bilinear interpolation used in other figures. These examples indicate that the different auxiliary classifiers generally agree on which are the most and least important parts of the image for the classification decision of the backbone, although the explanations produced by MobileNet-V3 and MnasNet are much coarser, due to their much lower resolution. This is in accord with the expected behavior since, in each case, the predictions of the same backbone are being explained. The explanation maps produced using the ResNet-50 auxiliary classifier are also coarser than those produced using ResNet-18, but more focused, explaining the moderate performance gains in Table 2. Finally, the explanation maps generated using the feature maps of ResNet-18 trained on Imagenette (denoted as "ResNet-18*") are noisier, with less well-ordered pixel importance, highlighting the importance of using a well-trained auxiliary classifier.

V. CONCLUSIONS

This paper presented P-TAME, a method for explaining DNN image classifiers by training an attention mechanism to combine feature maps produced by an auxiliary classifier into explanation maps, highlighting the important regions for the backbone model's prediction. P-TAME improves upon the paradigm established by T-TAME, extending it by decoupling the input of the attention mechanism responsible for producing explanations from the intermediate feature maps of the backbone being explained. This makes P-TAME a model-agnostic method, rendering it much more widely applicable. P-TAME produces explanation maps in a single forward pass during inference, while producing explanations that are on par with or better than those of the SoA explainability approaches. An important current limitation of P-TAME is the need for an auxiliary classifier trained on the backbone's original training data. A promising future direction is to investigate training the auxiliary classifier via knowledge distillation on the backbone itself, mitigating the need to procure the original training data and better tailoring it to the backbone being explained.

REFERENCES

[1] X. Li, H. Xiong *et al.*, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowl. Inf. Syst.*, vol. 64, no. 12, pp. 3197–3234, Dec. 2022.

[2] Z. Salahuddin, H. C. Woodruff *et al.*, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, p. 105111, Jan. 2022.

[3] G. Pavlidis, "Unlocking the black box: analysing the eu artificial intelligence act's framework for explainability in ai," *Law, Innovation and Technology*, vol. 16, no. 1, pp. 293–308, 2024.

[4] R. Saleem, B. Yuan *et al.*, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, vol. 513, pp. 165–180, Nov. 2022.

[5] S. Lapuschkin, S. Wäldchen *et al.*, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nat. Commun.*, vol. 10, no. 1, p. 1096, Mar. 2019.

[6] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[7] A. Chattopadhay, A. Sarkar *et al.*, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *Proc. IEEE WACV*, Mar. 2018, pp. 839–847.

[8] H. Wang, Z. Wang *et al.*, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," in *Proc. IEEE/CVF CVPR*, 2020, pp. 24–25.

[9] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized Input Sampling for Explanation of Black-box Models," in *British Machine Vision Conf. (BMVC)*, Sep. 2018.

[10] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization," in *Proc. IEEE WACV*, 2020, pp. 983–991.

[11] S. Sattarzadeh, M. Sudhakar *et al.*, "Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation," *Proc. AAAI*, vol. 35, no. 13, pp. 11 639–11 647, May 2021.

[12] M. Sudhakar, S. Sattarzadeh *et al.*, "Ada-Sise: Adaptive Semantic Input Sampling for Efficient Explanation of Convolutional Neural Networks," in *Proc. IEEE ICASSP*, Jun. 2021, pp. 1715–1719.

[13] B. Zhou, A. Khosla *et al.*, "Learning Deep Features for Discriminative Localization," in *Proc. IEEE/CVF CVPR*, 2016, pp. 2921–2929.

[14] M. Ntrougkas, N. Gkalelis, and V. Mezaris, "TAME: Attention Mechanism Based Feature Fusion for Generating Explanation Maps of Convolutional Neural Networks," in *IEEE Int. Symposium on Multimedia (ISM)*, Dec. 2022, pp. 58–65.

[15] M. V. Ntrougkas, N. Gkalelis, and V. Mezaris, "T-TAME: Trainable Attention Mechanism for Explaining Convolutional Networks and Vision Transformers," *IEEE Access*, vol. 12, pp. 76 880–76 900, 2024.

[16] J. Adebayo, J. Gilmer *et al.*, "Sanity Checks for Saliency Maps," in *Proc. NeurIPS*, Oct. 2018.

[17] A. E. Hassanien, D. Gupta *et al.*, *Explainable edge AI: a futuristic computing perspective*. Springer SCI vol. 1072, 2023.

[18] Y. Rong, T. Leemann *et al.*, "A Consistent and Efficient Evaluation Strategy for Attribution Methods," in *Proc. ICML*. PMLR, Jun. 2022, pp. 18 770–18 795.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., May 2015.

[20] K. He, X. Zhang *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE/CVF CVPR*, 2016, pp. 770–778.

[21] A. Dosovitskiy, L. Beyer *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, Oct. 2020.

[22] O. Russakovsky, J. Deng *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis. (IJCV)*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[23] H. Cappelen and J. Dever, *Making AI Intelligible: Philosophical Foundations*. Oxford University Press, 2021.

[24] A. Ghorbani, J. Wexler *et al.*, "Towards automatic concept-based explanations," in *Proc. NeurIPS*. Curran Associates Inc., Dec. 2019, no. 832, pp. 9277–9286.

[25] A. Sun, P. Ma *et al.*, "Explain Any Concept: Segment Anything Meets Concept-Based Explanation," in *Proc. NeurIPS*, Nov. 2023.

[26] A. Kumar, K. Sehgal *et al.*, "MACE: Model Agnostic Concept Extractor for Explaining Image Classification Networks," *IEEE Trans. Artif. Intell. (TAI)*, vol. 2, no. 6, pp. 574–583, Dec. 2021.

[27] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts," *Data Min. Knowl. Discov.*, vol. 38, no. 5, pp. 3043–3101, Sep. 2024.

[28] P. Coscia, A. Genovese *et al.*, "Features Disentanglement For Explainable Convolutional Neural Networks," in *IEEE Int. Conf. on Image Processing (ICIP)*, Oct. 2024, pp. 514–520.

[29] M. Caron, H. Touvron *et al.*, "Emerging Properties in Self-Supervised Vision Transformers," in *Proc. IEEE/CVF ICCV*, Oct. 2021, pp. 9630–9640.

[30] D. Zhou, Y. Shi *et al.*, "Refiner: Refining self-attention for vision transformers," *CoRR*, vol. abs/2106.03714, 2021.

[31] Z. Tu, H. Talebi *et al.*, "MaxViT: Multi-axis Vision Transformer," in *Proc. ECCV*. Springer Nature Switzerland, 2022, pp. 459–479.

[32] Z. Liu, Y. Lin *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *Proc. IEEE/CVF ICCV*. IEEE Computer Society, Oct. 2021, pp. 9992–10 002.

[33] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*. JMLR.org, Jun. 2013, pp. III–1310–III–1318.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proc. ACM SIGKDD*, Aug. 2016, pp. 1135–1144.

[35] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[36] H. Zhang, F. Torres *et al.*, "Opti-CAM: Optimizing saliency maps for interpretability," *Comput. Vis. Image Underst.*, vol. 248, p. 104101, Nov. 2024.

[37] X. Huang and J. Marques-Silva, "On the failings of Shapley values for explainability," *Int. J. Approx. Reason.*, vol. 171, p. 109112, Aug. 2024.

[38] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *Proc. IEEE/CVF CVPR*. IEEE, Jun. 2021, pp. 782–791.

[39] T.-Y. Lin, P. Dollár *et al.*, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE/CVF CVPR*, Jul. 2017, pp. 936–944.

[40] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[41] X. Cheng, Z. Rao *et al.*, "Explaining Knowledge Distillation by Quantifying the Knowledge," in *Proc. IEEE/CVF CVPR*, 2020, pp. 12 925–12 935.

[42] TorchVision maintainers and contributors, "Torchvision: Pytorch's computer vision library," https://github.com/pytorch/vision, 2016.

[43] A. Howard, M. Sandler *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF ICCV*. IEEE Computer Society, Oct. 2019, pp. 1314–1324.

[44] M. Tan, B. Chen *et al.*, "MnasNet: Platform-Aware Neural Architecture Search for Mobile," in *Proc. IEEE/CVF CVPR*. IEEE Computer Society, Jun. 2019, pp. 2815–2823.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proc. NeurIPS*, vol. 25. Curran Associates, Inc., 2012.

[46] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay," *CoRR*, vol. abs/1803.09820, 2018. [Online]. Available: http://arxiv.org/abs/1803.09820

[47] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, Sep. 2018.

[48] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. SPIE, May 2019, pp. 369–386.

[49] M. Balandat, B. Karrer *et al.*, "BOTORCH: A framework for efficient monte-carlo Bayesian optimization," in *Proc. NeurIPS*. Curran Associates Inc., Dec. 2020, pp. 21 524–21 538.

[50] V. C. Madala, S. Chandrasekaran, and J. Bunk, "CNNs Avoid the Curse of Dimensionality by Learning on Patches," *IEEE Open J. Signal Process. (OJSP)*, vol. 4, pp. 233–241, 2023.

[51] A. Hedström, L. Weber *et al.*, "A fresh look at sanity checks for saliency maps," in *World Conference on Explainable Artificial Intelligence*. Springer, 2024, pp. 403–420.

[52] J. Howard, "Imagenette: A smaller subset of 10 easily classified classes from imagenet," March 2019. [Online]. Available: https://github.com/fastai/imagenette

[53] P. Chowdhury, M. Prabhushankar *et al.*, "Are Objective Explanatory Evaluation Metrics Trustworthy? An Adversarial Analysis," in *IEEE Int. Conf. on Image Processing (ICIP)*, Oct. 2024, pp. 3938–3944.