FAME: Feature Activation Map Explanation on Image Classification and Face Recognition

Xinyi Zhang Manuel Günther
Department of Informatics, University of Zurich
xinyi.zhang@uzh.ch, siebenkopf@googlemail.com

Abstract

Deep Learning has revolutionized machine learning, reaching unprecedented levels of accuracy, but at the cost of reduced interpretability. Especially in image processing systems, deep networks transform local pixel information into more global concepts in a highly obscured manner. Explainable AI methods for image processing try to shed light on this issue by highlighting the regions of the image that are important for the prediction task. Among these, Class Activation Mapping (CAM) and its gradient-based variants compute attributions based on the feature map and upscale them to the image resolution, assuming that feature map locations are influenced only by underlying regions. Perturbation-based methods, such as CorrRISE, on the other hand, try to provide pixel-level attributions by perturbing the input with fixed patches and checking how the output of the network changes. In this work, we propose Feature Activation Map Explanation (FAME), which combines both worlds by using network gradients to compute changes to the input image, manipulating it in a gradient-driven way rather than using fixed patches. We apply this technique on two common tasks, image classification and face recognition, and show that CAM's above-mentioned assumption does not hold for deeper networks. We qualitatively and quantitively show that FAME produces attribution maps that are competitive state-of-the-art systems. Our code is available: https://github.com/AIML-IfI/fame.

1. Introduction

Image processing networks extract information from an image in highly non-linear and obscured ways to reach their excellent performance in several tasks, such as Image Classification (IC) and Face Recognition (FR). While early networks [15] were shallow enough for partial layer-wise visualization [35], modern architectures with hundreds of layers are far too complex to inspect individually. Therefore, research on eXplainable AI (XAI) has moved toward attribution methods, only highlighting the regions in the image that are influential to the decision process.

There exist two main streams of attribution methods. Gradient-based Class Activation Mapping (Grad-CAM) [30] and its variants make use of gradient information (see Sec. 3.1) and feature map activations to produce attribution maps. Since attributions are created in the size of the feature map, they are upscaled to image resolutions to provide pixel-level attributions. This upscaling is based on the assumption that elements in the feature map are influenced only by the underlying pixels. However, receptive fields of deep networks are much larger, and we show in our experiments that this assumption does not hold. Perturbation-based methods [13], on the other hand, do not make use of the network's gradient. Instead, they perturb local parts of the inputs and test how the prediction changes. For example, in an image classification task, when important parts of the input are perturbed, the prediction of the correct class should reduce [8]. However, the correct way to perturb images is unclear. For example, when adding local black squares or noise, such perturbations induce sharp edges into the input, which might represent features of some classes. When blurring local regions, this does not necessarily remove important information as modern networks are well-able to handle highly-blurred images [26].

In our proposed Feature Activation Map Explanation (FAME) we combine the two aspects by using network gradients that go back to the input level [12, 19] to compute attribution maps at the pixel level. Particularly, we exploit an iterative method that we had originally developed as adversarial attack [28], where we modify the original image in a more principled way in order to change network outputs, and interpret adversarial perturbations as an attribution map.

Depending on the target layer where we extract the network's output, and on the loss function used to obtain gradients, FAME can be applied to highlight various particularities. We start by extracting elements from the feature map and show that their receptive field is much larger for deeper networks than anticipated by CAM-based methods. Successively, we apply FAME on the predictions of classi-

fication networks, producing more fine-grained attribution maps that capture model-relevant sensitivity patterns beyond CAM-based methods. Finally, we show the generality of our method by highlighting similar and dissimilar regions of a pair of faces involved in a verification process. We quantitatively evaluate FAME against state-of-the-art methods and show that it is able to highlight important regions in the image similarly to or better than competitors. We also qualitatively show that FAME attributions cover more and better regions of objects than related methods.

2. Related Work

2.1. Deep Learning and the Black-Box Problem

Deep neural networks have achieved state-of-the-art performance across vision tasks such as Image Classifications (IC) [11] and Face Recognition (FR) [5]. In IC, networks f learn to hierarchically extract discriminative visual features φ and predict the logits z that assign semantic labels to entire images. In FR, given a pair of images, one from gallery x_p, and one from probe x_p, FR systems f learn to embed facial representations in a compact feature space, and compute the similarity to distinguish if they are from the same identity [25]. However, these deep networks are often considered as black boxes, making it difficult to understand the reasoning behind their predictions or detect potential biases and vulnerabilities [33]. This lack of transparency complicates the understanding of the performance degradations of IC and FR systems, which decrease under extreme conditions, such as variations in pose, quality, and occlusion.

2.2. Explainable AI techniques

EXplainable Artificial Intelligence (XAI) aims to make deep learning models more interpretable by addressing their black-box nature [17, 22]. Generally, XAI approaches are divided into intrinsic methods that design inherently interpretable models [29], and post-hoc methods that explain the behavior of complex pre-trained models without modifying their structure [16]. The latter have become the dominant choice for explaining deep neural networks due to their flexibility and scalability. Amongst these, Class Activation Mapping (CAM) [30] visualizes discriminative regions by backpropagating gradients from the output to feature maps. Its extensions, such as Grad-CAM++ [2] and HiResCAM [7], improve spatial precision. FullGradCAM [32] aggregates full-gradient contributions from both the input and all intermediate feature maps, thereby incorporating multi-layer information beyond single-layer approaches like Grad-CAM. Lately, research has turned to make use of gradients to the input [12, 19], see Sec. 3.2.

In contrast, perturbation-based methods generate explanations by measuring the impact of input modifications on model outputs. For example, Lu et al. [18] proposed CorrRISE to explain deep face verification models. They extended the RISE framework [24] to the face verification setting by computing Pearson correlation between randomly masked input face images with the similarity scores between two images. CorrRISE produces two saliency maps that highlight regions contributing to higher (*similar*) or lower (*dissimilar*) similarity. Although CorrRISE provides a baseline for explaining similarity decisions, it relies on random input masking, which introduces stochasticity and depends on manually-defined mask parameters.

2.3. Evaluation Metrics for XAI Techniques

Unfortunately, there is no universally-accepted metric for evaluating visual explanations, and different tasks often require different evaluation criteria. Adebayo et al. [1] even question whether using one AI model to validate another provides principled guarantees, showing that certain saliency maps are insensitive to model parameters. For IC, Intersection over Union (IoU) [36] is commonly used to measure the spatial overlap of attribution heatmaps with ground-truth regions, suitable for localized explanations. Remove and Debias (ROAD) [27] mitigates confounding effects in removal-based evaluations, improving reliability without requiring model retraining. Deletion and Insertion curves [24] evaluate how model confidence changes as salient pixels are progressively removed from the original image, or added to an empty canvas. Lu et al. [18] extended the Delete and Insert metric for face verification.

3. Method

Our explainable AI method, the Feature Activation Map Explanation (FAME) is built on the basis of our Layerwise Origin Target Synthesis (LOTS) [28] that we originally designed for adversarial attacks. LOTS iteratively computes pixel-level perturbations to the original image such that the network output at a certain layer represents a specific target. We reinterpret this method and turn it into an explainable AI technique, providing fine or coarse input-level attributions.

Subsequently, we make use of the following notation. For a given input image $x \in \mathbb{R}^{C\times H\times W}$ with C color channels and spatial dimensions $H \times W$, a deep network $f(x)$ is composed of $l = 1, \dots, L$ layers $f^l(\cdot)$ which are concatenated to produce O output logits $z \in \mathbb{R}^O$:

$$z = f^L\Big(f^{L-1}\Big(f^{L-2}\big(\cdots f^1(x)\cdots\big)\Big)\Big) \tag{1}$$

We particularly look into the output of two layers. One is the last convolutional layer $a = f^{l_a}(\cdot)$,[1] which is typically $l_a = L - 2$, that extracts the feature map $a \in \mathbb{R}^{C_a\times H_a\times W_a}$

[1] Similar concepts exist for Vision Transformer networks, but we leave the evaluation of these to future work.

with increased channels $C_a \gg C$ and reduced spatial dimensionality $H_a \times W_a \ll H \times W$. The other layer of interest is the embedding layer $\varphi = f^{l_\varphi}(\hat{a})$ (typically $l_\varphi = L - 1$) that processes a transformed feature map $\hat{a}$ into an embedding $\varphi \in \mathbb{R}^{C_\varphi}$. For IC, the feature map transform is often a global pooling with $\hat{a} \in \mathbb{R}^{C_a}$ and the logits z are typically used to predict the class of interest. For FR, the feature map a is flattened to $\hat{a} \in \mathbb{R}^{C_a \cdot H_a \cdot W_a}$ and specialized logits [6, 14, 21] are used to train discriminative features φ. Features φ_g and φ_p are extracted for a gallery x_g and a probe image x_p and compared using cosine, *i.e.*, the dot product of the normalized embeddings:

$$s = \frac{\varphi_g^\top \varphi_p}{\|\varphi_g\| \, \|\varphi_p\|} = \left(\frac{\varphi_g}{\|\varphi_g\|}\right)^\top \left(\frac{\varphi_p}{\|\varphi_p\|}\right). \tag{2}$$

3.1. Grad-CAM

Gradient-based Class Activation Mapping (Grad-CAM) makes use of the gradient [30]. In the case of image classification, this gradient is the partial derivative $\partial z_o / \partial a[k]$ of the logit for a given class $o \in [1, O]$ to a specific element $a[k] \in \mathbb{R}^{C_a}$ of the feature map with the location index $k \in [1, H_a] \times [1, W_a]$. These derivatives are combined with the feature map in various ways [2, 7] to produce an intermediate attribution map $\check{e} \in \mathbb{R}^{H_a \cdot W_a}$, which is then max-normalized and bilinearly interpolated to $e \in \mathbb{R}^{H \cdot W}$.

Grad-CAM was also extended to explain FR by backpropagating the similarity score s. Specifically, to avoid including the normalization factor from (2), Zhu et al. [37] compute $\partial \varphi_g^\top \varphi_p / \partial a_p[k]$ and $\partial \varphi_g^\top \varphi_p / \partial a_g[k]$.

3.2. FGGB

Recently, research has turned to using backpropagation of face embeddings φ to the input to compute attribution maps e. For example, Huber et al. [12] compute the Hadamard product between two embeddings: $v = \varphi_g \odot \varphi_p$, which they split into v_+ and v_-, depending on whether v_i exceeds a specific threshold θ. They backpropagate the average v_+ (and v_-) separately: $e_+ = \partial v_+ / \partial x_p$. Based on this, Lu et al. [19] proposed Feature-Guided Gradient Backpropagation (FGGB) by backpropagating each v_i to produce different e_i per embedding dimension i separately, and normalize the attribution maps individually. They compute weighted averages $e = \sum_i (v_i - \theta) e_i$, which can be split into similar and dissimilar maps e_+ and e_- by thresholding e at 0.

Both works make use of a fixed threshold θ that is selected based on the EER [12].[2] While FAME explores a similar idea, it does not rely on such an arbitrary threshold. Additionally, previous methods [12, 19] are constrained to a particular task, face verification, whereas FAME can be adapted to various attribution tasks.

[2] Actually, Lu et al. [19] does not explain how θ is selected, so we assume that this is identical to [12].

3.3. LOTS

The Layerwise Origin Target Synthesis (LOTS) developed by Rozsa et al. [28] is used to create an adversarial image $\bar{x} \in \mathbb{R}^{C \times H \times W}$ that fools the network to change its prediction toward a specific target t_l at layer l of the network. Starting from $\bar{x} = x$, LOTS makes use of the gradient of the loss $\mathcal{L}$, which is normalized by the maximum absolute value of the gradient, to update $\bar{x}$ iteratively:

$$\bar{x} \leftarrow \bar{x} - \eta \frac{\nabla_{\bar{x}}}{\max |\nabla_{\bar{x}}|} \qquad \nabla_{\bar{x}} = \frac{\partial \, \mathcal{L}\big(f^l(\bar{x}) \mid t^l\big)}{\partial \bar{x}}, \tag{3}$$

with a small step size $\eta > 0$. To create a FR attack, the adversarial probe image $\bar{x}_p$ is modified such that embedding $\bar{\varphi}_p = f^{l_\varphi}(\bar{x}_p)$ gets closer to the chosen gallery image $t_\varphi = f^{l_\varphi}(x_g) = \varphi_g$ using Euclidean loss $\mathcal{L}_2$ [28]. Importantly, LOTS uses the normalized raw gradient in (3), providing a different perturbation for each pixel, and not the sign as in the famous Fast Gradient Sign attack [9] and its variations.

3.4. FAME

Taking a closer look at the result of LOTS, the optimization procedure answers the question: *How do the pixels in the input image need to change to obtain the target as output?* This can be turned into the question: *Which pixels in the input image are most important to change the prediction?* by computing the adversarial perturbation:

$$\Delta x = \big|\bar{x} - x\big|, \tag{4}$$

and converting it to grayscale.[3] In order to turn this pixel-level annotation to a more broad region annotation, and in correspondence with [12, 19], the perturbation map Δx is then smoothed using a Gaussian kernel (see Sec. 4.4 for the effect of different blur kernel sizes), and max-normalized to provide an attribution map $e = [0, 1]^{H \times W}$. Through selecting specific target values t and loss functions $\mathcal{L}$, we apply FAME to produce attribution maps for different purposes:

FAME for Feature Maps. One important question that we ask in our evaluation is: *Which pixels from the input image contribute to single elements $a[k]$ in the feature map?* Such information is required to assess the viability of the assumption of Grad-CAM, *i.e.*, that only pixels under the feature map location influence the feature and, therefore, we can upscale the attribution map $\check{e}$ defined in Sec. 3.1. As the target t in (3), we select the zero vector. While this seems unintuitive upfront, we want to know: *Which image pixels have to be removed in order to extract no information at $a[k]$?* This is semantically similar to: *Which pixels contain information that influence $a[k]$?*

[3] Theoretically, we could also compute different attribution maps for the different color channels, but we leave this to future work.

For each feature map location k, we penalize only the activation at that spatial location and ignore all others:

$$\mathcal{L}_a(a[k] \mid 0) = \big\|a[k] - 0\big\|_1. \tag{5}$$

Applying FAME to $\mathcal{L}_a$ in (3) yields a local attribution at location k. Repeating this process over all locations k produces spatially resolved feature-attribution maps.

FAME for Image Classification. We follow the idea of Grad-CAM and define our loss as the logit of a given class o directly: $\mathcal{L}_{\text{cls}} = z_o$, which we backpropagate via (3). Unlike CAM-based methods that localize attribution regions in intermediate layers, FAME directly highlights input pixel.

FAME for Face Recognition. We make use of two normalized[4] embeddings φ_g and φ_p extracted from gallery x_g and probe image x_p, respectively, from which we calculate the similarity s via (2). Similar to Zhu et al. [37], we directly minimize the similarity itself to obtain the regions e_+ that are important in both images. Following [12, 18, 19], we estimate dissimilar regions e_- by minimizing $\mathcal{L}_-$, but without requiring any particular threshold:

$$\mathcal{L}_+(x_p \,|\, x_g) = s \qquad \mathcal{L}_-(x_p \,|\, x_g) = 1 - s\,. \tag{6}$$

When we backpropagate to either x_g or x_p, we keep φ_p or φ_g frozen, respectively. Intuitively, $\mathcal{L}_+$ highlights pixels whose modification most *reduces* similarity, *i.e.*, regions supporting the perceived similarity between the pair, whereas $\mathcal{L}_-$ emphasizes pixels whose change most *increases* similarity, revealing evidence for current dissimilarity. We conjecture, though, that defining dissimilar regions does not make sense in all cases; for example, it is difficult to manually analyze which regions in a non-matching pair should be most dissimilar. Thus, we show visual results for e_- as a reference, but focus our analysis on e_+.

4. Experiments

To demonstrate the versatility of our proposed FAME method across different visual understanding tasks, we conduct experiments on both IC and FR datasets. For IC, we employ the ImageNet validation set [4], which contains 50,000 images from 1,000 object categories, providing diverse visual concepts and well-defined bounding boxes, allowing quantitative evaluation of attribution maps. In our experiments, we select 5 images for each of the 1000 classes that are correctly classified by all employed networks, and back-propagate the logits z_o of the ground truth class.

For the FR task, we use three representative datasets: AR Face [20], CFP [31], and SCface [10], since they allow for an unadulterated evaluation of specific conditions of occlusions, face pose and image resolution, in opposition to other benchmark datasets that mix these conditions [18]. The AR Face dataset contains variations in occlusions caused by scarves covering the mouth and sunglasses covering both eyes, and we follow standard protocols [3]. The CFP dataset consists of frontal and profile faces with yaw angles greater than 60°, where only one side of the face is visible, and we adopt the default protocols [31]. SCface is captured by indoor surveillance cameras under different distances and resolutions, where the default protocols [34] allow systematic evaluation of distance effects. We follow standard image preprocessing and facial alignment routines.

During experiments, we compare performance of different models. For IC, we test three pretrained models: ResNet34, ResNet50, and ResNet101, as well as the old VGG19 network and the state-of-the-art ConvNeXt-Tiny variant from the PyTorch [23] model zoo. For FR, to analyze the impact of network depth on interpretability we adopt IResNet101[5] along with two smaller variants, IResNet18 and IResNet50 [14], all pre-trained with AdaFace.

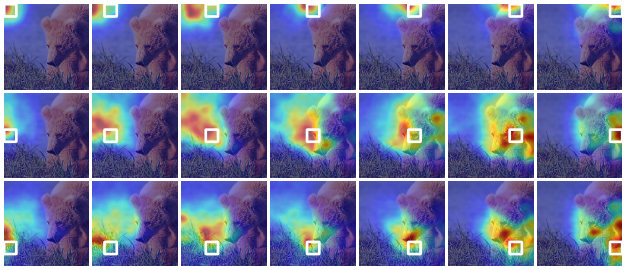
Moreover, we compare our method with several representative XAI techniques, including Grad-CAM [30] and Grad-CAM-Elementwise (Grad-CAM-EW). For IC, we additionally compare against HiResCAM [7] and FullGrad-CAM [32], for FR we add CorrRISE [18] and FGGB [19], to demonstrate the effectiveness and generalizability of FAME across both tasks. For CorrRISE [18], we follow most of the implementation details described in their paper. Specifically, 500 random masks are generated for each image pair, where each mask contains 10 black patches with the size of 30×30 that are multiplied with the original gallery and probe images to replace the corresponding pixels separately. Since the official implementation is not publicly available, we reproduce FGGB strictly following the procedural details described in the original paper [19]. For FAME, we rely on the original LOTS parameters [28], *i.e.*, we use $\eta = 1/255$ and 500 iterations. For computing the FAME attribution, we use a Gaussian blur with standard deviation of 7.7, see Sec. 4.4 for a discussion on this choice.

4.1. How Reliable is CAM?

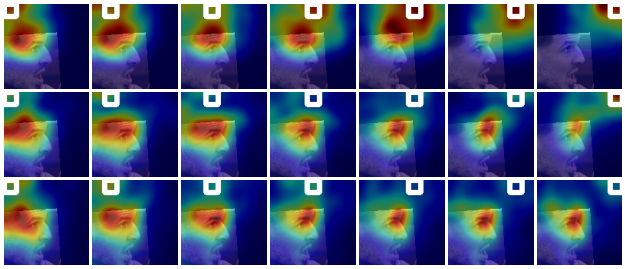
As stated above, CAM-based visualization methods assume that elements in the feature map are only influenced by underlying pixel areas. In this section, we use FAME via (5) to falsify this assumption. Fig. 1(a) reveals that different spatial positions in the feature map correspond to regions of varying size and coverage in the input image. While some locations focus on localized object parts (*e.g.*, the bear's head or ear), others capture information from much broader areas, extending across the body and surrounding background. This observation confirms that the receptive fields of feature-map elements are neither uniform nor localized. Some units integrate context from large regions of the input image rather than from spatially fixed patches. Thus, inter-

[4] We do not compute gradients for the denominators in (2).

[5] https://github.com/mk-minchul/AdaFace



(a) Image Classification



(b) Face Recognition

Figure 1: FEATURE MAP VISUALIZATION. The figure shows partial feature-map visualizations obtained using the proposed FAME method via $\mathcal{L}_a$ on the 7×7 feature map a. In (a), we show one exemplary ImageNet image, for rows 1, 4 and 6 of the feature map using ResNet101. In (b), we show the first row of features on a profile face image, using three networks IResNet34, IResNet50, and IResNet101. White borders indicate the pixels that lie underneath the particular feature map location $a[k]$.

preting CNNs through upsampled activation-level maps, as in CAM-based methods, can lead to misleading attribution.

To further verify this phenomenon, we apply FAME on FR networks, where we systematically examine how feature-map activations correspond to different input regions across architectures of varying depth. As shown in Fig. 1(b), the shallow network (IResNet34) tends to extract information from relatively fixed local areas, *i.e.*, each spatial unit in the feature map corresponds to a specific region in the input. In contrast, receptive fields of deeper networks (IResNet50 and IResNet101) expand and shift, causing activations at the top positions of the feature map to aggregate information from central regions of the input face rather than spatially aligned areas. FAME provides more reliable spatial interpretation, since the elements of feature maps are not fixed representations of input regions, especially in deeper networks where receptive fields become large and overlapping. It demonstrates that deeper architectures encode more abstract, globally integrated patterns, flagging CAM-based methods as unreliable.

4.2. Image Classification

Visual Results. Fig. 2 qualitatively compares saliency maps generated by Grad-CAM-EW and FAME across different network architectures, including ResNet variants, VGG19, and ConvNeXt-Tiny, on two ImageNet samples. For the bird, Grad-CAM-EW often produces diffuse responses that extend to the surrounding context, such as grass, or concentrates on the bird's head only, whereas FAME highlights regions concentrated around visually salient locations across the bird, across different backbones. In the elephant example, Grad-CAM-EW typically activates broad areas mostly covering the head, while FAME emphasizes more compact regions around the entire body. Notably, attribution patterns of FAME are similar across all networks. These qualitative results indicate that FAME can be applied consistently across diverse deep network architectures, producing comparable sensitivity patterns.

Table 1: QUANTITATIVE EVALUATION OF ATTRIBUTIONS FOR IMAGE CLASSIFICATION. This table shows the quantitative comparison of different XAI methods on ImageNet. Higher IoU indicates better spatial alignment with ground-truth object regions, while higher ROAD-Delete scores indicate faster performance degradation when salient regions are removed.

	IoU in % ↑			ROAD-Delete ↑		
XAI	thr=0.3	thr=0.5	thr=0.7	P=10%	P=30%	P=50%
Grad-CAM	38.76	23.38	10.24	*0.2665*	*1.0354*	2.2562
Grad-CAM-EW	40.15	24.40	10.70	0.2518	0.9749	2.1243
HiResCAM	38.76	23.38	10.24	*0.2665*	1.0354	*2.2563*
FullGradCAM	**48.81**	**33.05**	**13.14**	0.2527	1.0200	**2.2888**
FAME - $\mathcal{L}_{\text{cls}}$	*46.09*	*29.09*	*10.79*	**0.4253**	**1.1499**	2.1672

Evaluation Metrics. To assess the quality of the explanations produced by FAME, we employ two complementary evaluation metrics. Particularly, we adopt the Intersection over Union (IoU) metric as a localization proxy to quantify the spatial overlap between attribution maps and ground-truth object regions. While IoU does not constitute a complete measure of explanation faithfulness, it remains a widely used evaluation metric in recent work [16]. We report IoU at multiple thresholds ($\texttt{thr} = 0.3, 0.5, 0.7$), where lower thresholds yield broader saliency and higher thresholds produce sparser, more discriminative maps. This evaluation allows us to analyze the robustness of different XAI methods under varying levels of strictness in saliency selection. Complementing IoU, we adopt ROAD-Delete [27] as a removal-based faithfulness proxy, where removal ratio P denotes the percentage of pixels removed according to the attribution ranking. It evaluates how rapidly model performance degrades when the most salient regions are removed. Consistent performance degradation across different P values indicates that the explanation reliably captures regions that are critical to the model's decision. While ROAD-Insert is used for our FR part later, we do not adopt it for IC, since object classification requires larger connected regions, whereas pixel-level attribution methods produce more localized and disjoint regions in IC (*cf*. Fig. 2), while FR re-

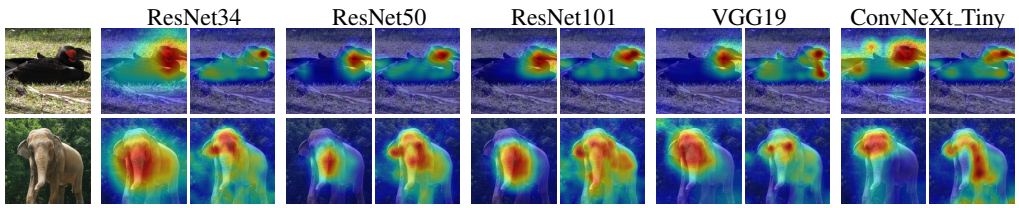
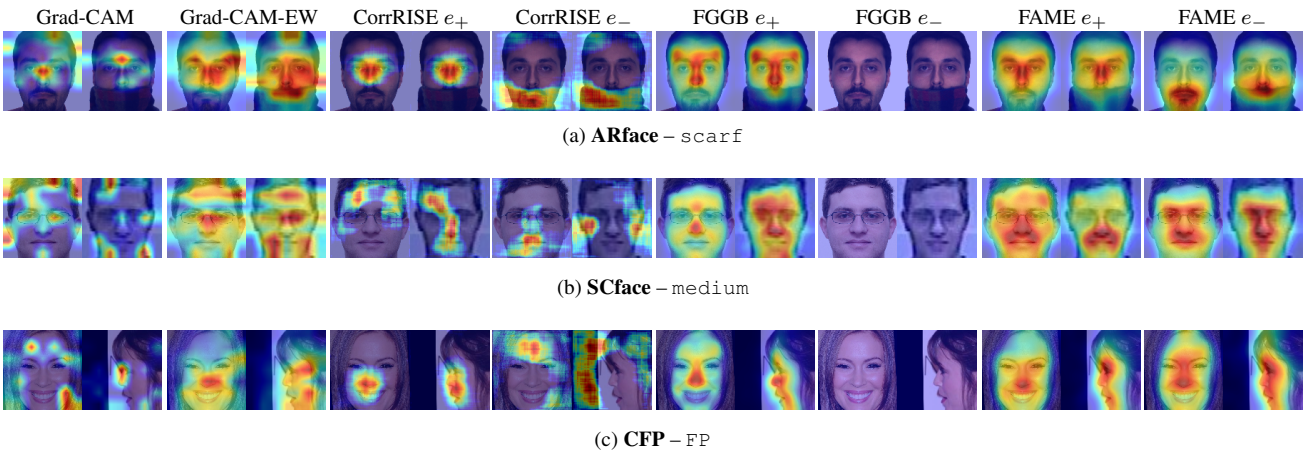


Figure 2: IMAGENET VISUALIZATION. The figure shows the saliency maps generated by Grad-CAM-EW (left in each pair) and FAME (right) using different models on two ImageNet samples, evaluated with five different pre-trained networks.



(a) **ARface** – `scarf`

(b) **SCface** – `medium`

(c) **CFP** – `FP`

Figure 3: EXPLAINING FACE RECOGNITION. The figure provides a comparative visualization of explanation maps generated on three representative datasets for genuine (same-identity) image pairs using IResNet101. XAI techniques include Grad-CAM, Grad-CAM-EW, CorrRISE, FGGB, and FAME, including similar e_+ and dissimilar attribution e_- where appropriate.

lies on more localized features.

Quantitative Evaluation. Tab. 1 reports the quantitative comparison of different attribution methods on ImageNet. In terms of IoU, FAME achieves competitive localization performance compared to CAM-based methods while remaining slightly below FullGradCAM across different thresholds. This indicates that FAME provides reasonable spatial alignment with object regions. However, not all parts of the object are important for classification, which shows conceptual limitations of the IoU metric for evaluation attribution. When using ROAD-Delete to evaluate explanation quality through removal-based performance degradation, FAME achieves the highest scores at $P = 10\%$ and 30%, and remains competitive at higher removal ratios. This suggests that the most important regions highlighted by FAME are particularly relevant to the model prediction.

4.3. Face Recognition

Visual Results. Fig. 3 compares attribution maps from different XAI methods on three representative datasets using the largest IResNet101 model; additional backbones are shown in the supplemental material. Grad-CAM averages channel weights across spatial locations (*cf*. [30]), which assumes feature map pooling. Since FR networks do not perform such pooling (*cf*. Sec. 3), the resulting attributions appear spatially inconsistent. Grad-CAM-EW highlights non-specific and spatially disjoint facial regions. CorrRISE produces separated regions for e_+, while e_- often corresponds to background. These patterns remain noisy, particularly for small or low-resolution inputs in Fig. 3(b). FGGB decomposes the evidence into two components based on the threshold θ (cf. Sec. 3.2), selected at EER [12, 19]. Although it highlights e_+ reasonably well, the spatial patterns are less stable across datasets, especially under occlusion or resolution degradation. The largely missing e_- visualization in Fig. 3 reflects its sensitivity to the threshold θ.

FAME produces more spatially coherent and consistent attribution patterns across datasets. On ARFace, FAME's e_+ concentrates on non-occluded identity cues such as the eyes and forehead, ignoring the covered mouth region, whereas e_- highlights this region — for both gallery and probe images. For SCFace, FAME remains robust to resolution degradation, with central activation near the nose and mouth. On CFP, FAME captures the correspondence between frontal and profile views, focusing on shared structural areas like the nose bridge and cheek contour. Interestingly, this holds for both the similar and dissimilar maps,

Table 2: QUANTITATIVE EVALUATION OF FACE RECOGNITION ATTRIBUTION. This table compares different XAI methods evaluated using the Delete (↓) and Insert (↑) metrics across multiple FR protocols. Three AdaFace-based backbones (IResNet18, IResNet50, IResNet101) are tested on three datasets: AR Face, SCface, and CFP with default evaluation protocols. **Best** and *second-best* results per model, protocol, and metric are highlighted.

FR model	XAI	ARface						SCface						CFP			
		neutral		glass		scarf		close		medium		far		FF		FP	
		Delete	Insert	Delete	Insert	Delete	Insert	Delete	Insert	Delete	Insert	Delete	Insert	Delete	Insert	Delete	Insert
IResNet18	Grad-CAM	*50.51*	50.32	63.77	65.63	61.71	81.78	66.56	70.33	62.47	63.03	53.64	55.73	76.94	89.02	55.04	66.79
	Grad-CAM-EW	**50.47**	52.48	54.35	*84.80*	57.42	90.96	58.17	*85.90*	55.95	**78.74**	51.88	**60.90**	64.51	*95.45*	52.31	*78.54*
	CorrRISE e_+	**50.47**	50.00	53.69	80.20	*53.84*	*92.66*	*55.87*	84.23	*54.59*	74.17	**51.30**	59.56	69.19	94.43	*51.94*	77.90
	FGGB e_+	**50.47**	*52.69*	**52.85**	75.93	54.10	85.44	**55.83**	80.13	**53.74**	72.91	*51.35*	58.91	*62.89*	88.10	52.19	73.31
	FAME e_+	**50.47**	**52.78**	*53.06*	**85.43**	**53.41**	**92.86**	56.80	**86.22**	55.05	*78.53*	52.40	*60.67*	**60.75**	**95.58**	**51.87**	**79.96**
IResNet50	Grad-CAM	*50.90*	50.81	75.23	75.21	68.85	78.29	69.62	75.81	63.00	67.12	54.63	57.34	79.11	88.60	70.09	83.29
	Grad-CAM-EW	**50.77**	*54.23*	59.71	90.88	57.44	91.20	60.48	87.92	56.87	*78.17*	52.64	*62.71*	65.65	95.48	56.78	93.61
	CorrRISE e_+	**50.77**	50.00	59.55	*91.18*	*55.07*	*92.64*	60.07	*88.62*	*56.30*	76.53	**52.06**	60.20	72.68	**96.27**	*55.69*	*94.68*
	FGGB e_+	**50.77**	54.02	*57.34*	82.54	57.40	85.11	*59.59*	82.78	56.48	72.01	*52.31*	59.97	*64.78*	88.36	57.55	88.91
	FAME e_+	**50.77**	**55.00**	**56.85**	**91.75**	**54.46**	**92.84**	**58.31**	**88.94**	**56.17**	**78.97**	53.43	**63.14**	**62.20**	*96.00*	**54.84**	**95.05**
IResNet101	Grad-CAM	*50.64*	50.60	72.62	77.89	69.48	80.50	71.27	75.40	67.64	69.86	54.51	54.40	86.41	91.94	74.91	85.21
	Grad-CAM-EW	**50.60**	52.91	60.44	90.09	59.12	*90.80*	60.57	*87.17*	59.05	*80.56*	52.41	*59.48*	70.17	96.70	59.26	94.36
	CorrRISE e_+	**50.60**	50.00	58.39	*91.24*	*55.52*	**93.43**	*58.59*	86.78	*57.35*	80.35	**51.38**	58.79	86.46	**96.81**	*56.31*	*96.07*
	FGGB e_+	**50.60**	*53.63*	*58.09*	83.71	58.17	86.61	59.59	83.17	58.49	75.58	*52.17*	57.91	*69.10*	91.39	58.66	90.85
	FAME e_+	**50.60**	**54.06**	**56.12**	**92.07**	**55.13**	**93.43**	**58.10**	**88.22**	**57.19**	**82.02**	52.73	**60.15**	**65.70**	*96.71*	**55.82**	**96.24**

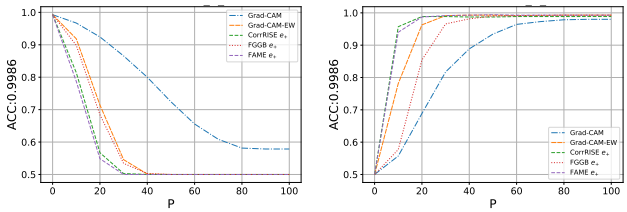
which better follow our intuition than CorrRISE,[6] which arbitrarily highlights forehead or background regions.

Evaluation Metrics. We follow the *Deletion* and *Insertion* strategy [18], which measures how verification accuracy changes as salient pixels are progressively removed or inserted. For each image pair, we compute the similarity score s via (2). Given an attribution map e_+, we sort pixels by importance and iteratively delete or insert the top-P percent ($P \in \{0, 10, 20, \ldots, 100\}$) on the probe image x_p, while keeping the gallery image x_g fixed to isolate the attribution effect. After each perturbation, we recompute similarity and evaluate verification accuracy using the original decision threshold, yielding an accuracy-over-P curve. We summarize performance by the normalized area under this curve (AUC). A faithful explanation produces a steeper accuracy drop (lower AUC) under deletion and a faster recovery (higher AUC) under insertion.

Quantitative Evaluation. Tab. 2 presents the quantitative comparison of different XAI methods evaluated using the Delete and Insert metrics across multiple FR protocols. Overall, FAME consistently achieves competitive or superior scores compared with CAM-based approaches (Grad-CAM, Grad-CAM-EW), the perturbation-based CorrRISE, and the state-of-the-art FGGB. In most protocols, FAME yields higher Insert values and comparable or lower Delete values, indicating that the regions identified by FAME are both highly informative and strongly aligned with the model's similarity computation.

On ARFace, FAME particularly improves under occluded conditions such as a glass and a scarf, suggesting better robustness in identifying the true discriminative regions when facial features are partially covered. Similar improvements are observed on SCface, where FAME demonstrates more stable performance across varying camera distances, reflecting stronger generalization. On CFP, FAME achieves the highest Insert scores (up to 96.7%), confirming its effectiveness under pose variations. Compared to CAM-based methods, which are often affected by feature-map misalignment, cf. Sec. 4.1, FAME provides more consistent activation relevance by directly modeling the relationship between input regions and the similarity.

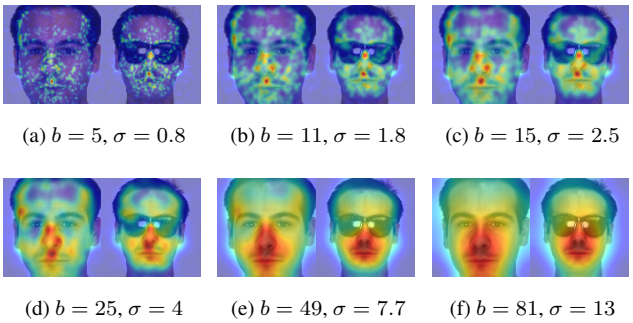
Fig. 4 compares the Deletion and Insertion curves of different XAI methods on the CFP FP protocol using the IResNet101 model. In Fig. 4(a), accuracy decreases as the most salient pixels are progressively removed from the probe image, while in the Fig. 4(b), accuracy increases as pixels are gradually added back. As shown, FAME consistently out-



(a) Deletion (b) Insertion

Figure 4: DELETION VS. INSERTION. This figure shows (a) Deletion and (b) Insertion evaluation curves for different XAI methods on the CFP FP protocol using IResNet101, which has a clean accuracy of 99.86 %.

[6] It is difficult to assess which facial regions are dissimilar. However, highlighting background that does not influence embeddings is incorrect.



(a) $b = 5, \sigma = 0.8$ (b) $b = 11, \sigma = 1.8$ (c) $b = 15, \sigma = 2.5$

(d) $b = 25, \sigma = 4$ (e) $b = 49, \sigma = 7.7$ (f) $b = 81, \sigma = 13$

Figure 5: EFFECT OF GAUSSIAN SMOOTHING. This figure shows attribution maps produced by the proposed FAME method under different Gaussian blur parameters, where b denotes the blur kernel size in pixels, and σ the standard deviation of the Gaussian.

performs CAM-based methods, exhibiting the fastest performance degradation under Deletion and the fastest accuracy gain by Insertion. While CorrRISE has a slightly steeper increase for small P in Fig. 4(b), indicating the most important regions are highlighted better, FAME surpasses CorrRISE at higher P, showing that it captures all relevant pixels more reliably. This demonstrates that FAME's attribution maps identify regions that are highly relevant to face verification. These quantitative results, together with the qualitative analysis in Fig. 3, show that FAME produces visually meaningful explanations and better correlates with similarity scores across network depths and datasets.

4.4. Effect of Gaussian Blur

To visualize the effect of different Gaussian blur sizes, we evaluate the IResNet101 model on one ARFace sample. As shown in Fig. 5, the qualitative structure of the attribution maps evolves smoothly with the degree of smoothing. When the kernel size is small, as in Fig. 5(a), the maps retain fine-grained, high-frequency responses reflecting the raw perturbation field. Although such maps appear noisy, the central activations are around the nose tip and mouth region. As the blur increases to moderate values in Fig. 5(c), the explanations become spatially coherent while preserving the main evidence regions. This setting provides a good balance between visual smoothness and localization fidelity: the highlighted areas correspond to key identity cues (nose, mouth, and cheek contours) while suppressing noisy fluctuations and irrelevant activations on the forehead or chin. In contrast, large smoothing in Fig. 5(f) produces overly diffuse maps where almost the entire face is highlighted, leading to loss of spatial contrast and interpretive precision.

Overall, the results demonstrate that Gaussian smoothing acts as an effective regularizer for FAME's perturbation field, suppressing artifacts without altering the fundamental explanatory structure. The persistence of nose- and mouth-centered activations across all parameter choices further confirms that FAME captures robust and semantically meaningful evidence for identity similarity even under occlusion conditions, such as with sunglasses. For increased comparability to CAM-based attribution maps, and consistent with [12, 19], we used a blur with a kernel size of 7.7 as shown in Fig. 5(e). This seems appropriate to the two tasks that we perform, *i.e.*, IC on ImageNet, and FR. However, other tasks might require more fine-grained visualization, for which smaller kernel sizes would be more appropriate.

5. Conclusion

In this work, we proposed our Feature Attribution Map Explanation (FAME) technique, a unified explainability framework that bridges gradient-based and perturbation-based paradigms for visual interpretation of deep neural networks. Unlike conventional CAM-based methods that rely on a fixed spatial correspondence between feature maps and input regions, FAME explicitly models the pixel-level influence on intermediate activations through controlled gradient-driven perturbations. It also removes the necessity to define thresholds as required by the most related FGGB method. Comprehensive experiments on both IC and FR tasks demonstrate the versatility and effectiveness of the method, which works across various network topologies, including those performing pooling and flattening of the feature maps. On ImageNet, FAME achieves competitive localization performance as measured by IoU, while emphasizing regions that are more relevant under deletion-based evaluation beyond the current state-of-the-art Full-GradCAM. On multiple FR datasets, FAME consistently outperforms Grad-CAM-EW, HiResCAM, CorrRISE and FGGB under both qualitative and quantitative evaluations, showing superior interpretation of occlusion, pose, and resolution variations, without requiring to select any threshold parameters to separate similar and dissimilar attribution maps. Furthermore, our analysis on feature maps reveals that deeper networks capture information from increasingly large and shifted receptive fields, confirming that naïve spatial upsampling in CAM-based techniques is unreliable.

Overall, FAME provides a task-adaptive, optimization-based framework that yields stable and model-relevant sensitivity-based explanations across IC and FR in our experiments. Future work will extend FAME to transformer-based architectures and other tasks, such as object detection, and explore its integration into human-centered explainability frameworks for trustworthy AI systems. Notably, FAME is relatively slow due to the iterative gradient steps used to compute the adversarial sample via LOTS. We used the default parameters proposed by Rozsa et al. [28]. The supplemental material shows that larger step sizes η, fewer iterations, and loss-specific early stopping can increase speed with little impact on visualization quality.

Acknowledgements

This work is supported by the University of Zurich via the UZH Candoc Grant, grant no. FK-25-016.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[2] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018. 2, 3

[3] Tiago de Freitas Pereira, Dominic Schmidli, Yu Linghu, Xinyi Zhang, Sébastien Marcel, and Manuel Günther. Eight years of face recognition research: Reproducibility, achievements and open issues. *arXiv*, 2022. 4

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009. 4

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[6] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(10):5962–5979, 2022. 3

[7] Rachel Lea Draelos and Lawrence Carin. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv*, 2020. 2, 3, 4

[8] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representation (ICLR)*, 2015. 3

[10] Mislav Grgic, Kresimir Delac, and Sonja Grgic. SCface - surveillance cameras face database. *Multimedia Tools and Applications*, 51(3), 2011. 4

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016. 2

[12] Marco Huber, Anh Thi Luu, Philipp Terhörst, and Naser Damer. Efficient explainable face verification based on similarity score argument backpropagation. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1, 2, 3, 4, 6, 8

[13] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021. 1

[14] Minchul Kim, Anil K. Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012. 1

[16] Xingjian Li, Qiming Zhao, Neelesh Bisht, Mostofa Rafid Uddin, Jin Yu Kim, Bryan Zhang, and Min Xu. DiffCAM: Data-driven saliency maps by capturing feature differences. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 5

[17] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106, 2024. 2

[18] Yuhang Lu, Zewei Xu, and Touradj Ebrahimi. Towards visual saliency explanations of face verification. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2, 4, 7

[19] Yuhang Lu, Zewei Xu, and Touradj Ebrahimi. Explainable face verification via feature-guided gradient backpropagation. In *International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024. 1, 2, 3, 4, 6, 8

[20] Aleix M. Martínez and Robert Benavente. The AR face database. Technical Report 24, Universitat Autònoma de Barcelona, CVC, 1998. 4

[21] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[22] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 2023. 2

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2019. 4

[24] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018. 2

[25] P. Jonathon Phillips, Patrick Grother, and Ross Micheals. *Handbook of Face Recognition*, chapter Evaluation Methods in Face Recognition. Springer, 2nd edition, 2011. 2

[26] Wes Robbins, Gabriel Bertocco, and Terrance E. Boult. DaliID: Distortion-adaptive learned invariance for identification – a robust technique for face recognition and person re-identification. *IEEE Access*, 2024. 1

[27] Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning (ICML)*. PMLR, 2022. 2, 5

[28] Andras Rozsa, Manuel Günther, and Terranee E. Boult. LOTS about attacking deep features. In *International Joint Conference on Biometrics (IJCB)*. IEEE, 2017. 1, 2, 3, 4, 8

[29] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 2018. 2

[30] 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 *International Conference on Computer Vision (ICCV)*. IEEE, 2017. 1, 2, 3, 4, 6

[31] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *Winter Conference on Applications of Computer Vision (WACV)*, 2016. 4

[32] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 4

[33] Cesare Tucci, Attilio Della Greca, Genoveffa Tortora, and Rita Francese. Explainable biometrics: A systematic literature review. *Journal of Ambient Intelligence and Humanized Computing*, 2024. 2

[34] Roy Wallace, Mitchell McLaren, Chris McCool, and Sébastien Marcel. Cross-pollination of normalisation techniques from speaker to face authentication using Gaussian mixture models. *Transactions on Information Forensics and Security (TIFS)*, 7(2), 2012. 4

[35] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014. 1

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[37] Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual explanation for deep metric learning. *Transactions on Image Processing (TIP)*, 30:7593–7607, 2021. 3, 4