Faces in Focus: Evaluating Generalizability and Transferability of Facial Recognition Explainability Techniques for CNN-based Models

Paweł Borsukiewicz[1], El-hacen Diallo[1], Abdoul Aziz Bonkoungou[1], Wendkûuni C. Ouédraogo[1], Jacques Klein[1], Charles Beumier[2], Tegawendé F. Bissyandé[1]

[1] University of Luxembourg, Luxembourg
[2] Royal Military Academy, Belgium

{pawel.borsukiewicz, el-hacen.diallo, abdoul.bonkoungou, wendkuuni.ouedraogo, jacques.klein, tegawende.bissyande}@uni.lu charles.beumier@mil.be

Abstract

Most explainability (XAI) methods for facial recognition provide pairwise explanations, highlighting why two specific images match. Privacy-oriented applications such as face de-identification, however, need to identify regions that are critical to an individual's identity across multiple images. This calls for XAI methods that generalize (reliably locating identity-critical regions across different images) and that transfer (producing regions on one model that remain meaningful for others). We benchmark two families of XAI approaches, activation-mapping-based (Grad-CAM, Grad-CAM++, LayerCAM) and perturbation-based (S-RISE, CorrRISE), on their ability to localize identity-critical facial regions. Our evaluation covers seven CNN architectures and three datasets under both white-box and black-box settings. To measure identity-level importance, we apply five occlusion protocols that test whether masking highlighted regions disrupts recognition of the target identity. Grad-CAM++ most consistently identifies identity-critical regions across models and datasets. S-RISE, CorrRISE, and LayerCAM frequently highlight areas with little impact on recognition, while Grad-CAM proves reliable only for a narrow subset of models. We also provide practical guidance on target-layer selection and loss-function design, two factors we find to be key drivers of identity-level generalizability. Finally, we observe strong cross-model transferability of the identified regions, a property particularly valuable for models trained on synthetic data, where XAI methods tend to underperform.

1. Introduction

Explainable AI (XAI) makes model decisions understandable by revealing the evidence behind a prediction. Beyond improving transparency, XAI also supports accountability, fairness, and compliance with emerging AI regulations such as the right to explanation, especially for high-risk systems [14]. In computer vision, neural network (NN) explainability is widely used to interpret model outputs [19, 38]. By analyzing model focus, we can quantify how strongly image regions contribute to the output [46]. These insights enable debugging, redesign, and audits [31].

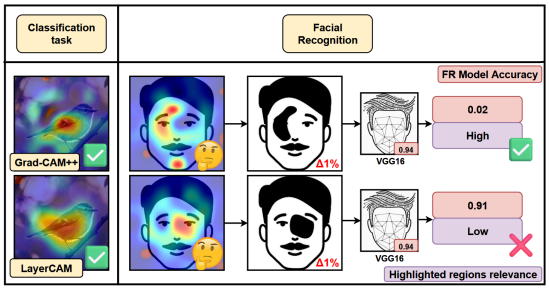

Figure 1. XAI evaluation is simple in classification, but facial recognition lacks ground-truth identity regions and requires different evaluation criteria.

Most NN explainability techniques in computer vision are defined for classification, where heatmaps indicate regions that raise specific class scores [12]. Face recognition (FR) is different: decisions are made from distances or similarities in an embedding space, so a saliency map must be read as highlighting regions that increase (positive evidence) or decrease (negative evidence) the similarity for a given identity claim [22, 24]. Unlike object localization (Figure 1), FR has no pixel-level ground truth for "identity-critical" regions. Therefore, faithfulness should instead be judged by counterfactual effect: if the highlighted pixels are hidden, does the verification score change as predicted?

Recent studies have evaluated FR explainability using pairwise insertion and deletion metrics, reporting

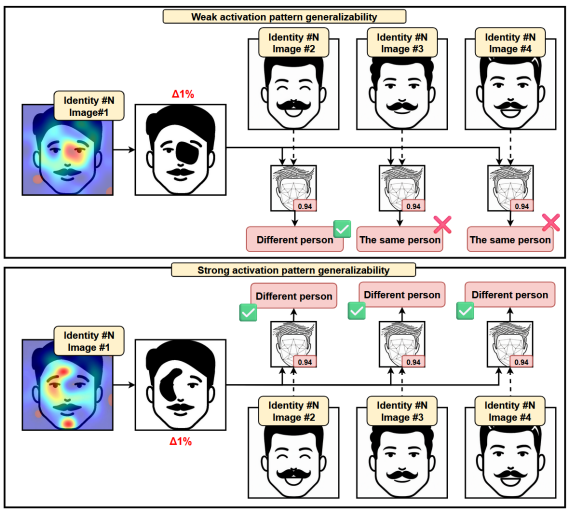

Figure 2. Generalizability refers to the identified facial regions being relevant to a broader range of images.

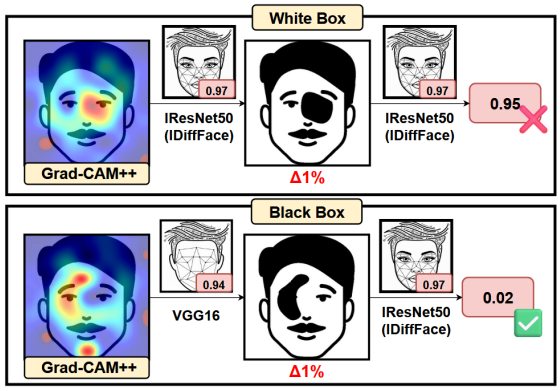

Figure 3. Transferability refers to successful reuse of key facial regions, as indicated by a source model, to impact a target model.

strong performance for S-RISE [24] and positioning CorrRISE [25] above Grad-CAM family methods such as Grad-CAM [33] and Grad-CAM++ [10]. However, these evaluations are pairwise and do not test whether the highlighted regions generalize across different images of the same person. This limitation matters in privacy-preserving scenarios, where an adversary may possess multiple photos of a given identity: a pattern tuned to one picture can fail on others. We address this gap by studying generalizability (Figure 2) and cross-model transferability (Figure 3) of FR explainability techniques. We consider three CAM methods (Grad-CAM, Grad-CAM++, LayerCAM [21]) and two perturbation methods (S-RISE, CorrRISE) at the identity-level: we select salient regions for each individual and measure their impact on verification across multiple images of that person, rather than on a specific image pair as in CorrRISE [25]. The contributions of this paper are summarized as follows:

- We study the generalizability and transferability of two families of explainability methods for facial recognition across seven CNN-based models and three datasets.
- We adapt these methods and determine the optimal hyper-parameters for models trained on both real and synthetic data.
- We show that Grad-CAM++ is the only method that reliably highlights identity-critical regions; LayerCAM, S-RISE, and CorrRISE mainly target areas with negligible impact on recognition.
- We report a discrepancy, with XAI techniques proving less effective for models trained on synthetic vs. real data.
- We demonstrate cross-model transferability of the most relevant areas.

Supplementary material with more in-depth analyses is available here[1].

2. Related Work

2.1. Image Recognition Explainability

Explainable image recognition techniques identify regions that influence the model prediction. Various approaches have been proposed to meet this objective [23, 34].

One category includes variants of class activation mapping (CAM). The first CAM method was introduced by Zhou et al. [46] over a decade ago. Since then, several improved variants have been proposed. Grad-CAM [33] was designed to overcome the original CAM constraint that required a global average pooling layer before classification. It combines activation values from the forward pass with gradients from the backward pass to estimate region relevance. Because Grad-CAM relies on first-order derivatives and produces one gradient weight per channel, it struggles to capture fine-grained, pixel-level details. Grad-CAM++ [10] addresses this limitation by using second- and third-order derivatives to compute pixel-dependent weights. However, the aforementioned CAM-based methods were typically applied to a final convolutional layer. Because of their more class-discriminative features CAM-based methods still cannot be applied uniformly to arbitrary convolutional layers. LayerCAM [21] introduced pixel-specific gradients to overcome this restriction and improve localization across different CNN layers.

The second category of explanation methods relies on adversarial perturbations, such as black-pixel masking, and evaluates the resulting drop in accuracy. LIME [31] was introduced as a model-agnostic technique that estimates feature relevance using outputs from slightly perturbed images. Problems with inconsistent explanations have been addressed by its extensions, including OptiLIME [37], BayLIME [43] and S-LIME [47].

[1] https://github.com/Pabilito/Faces-in-Focus

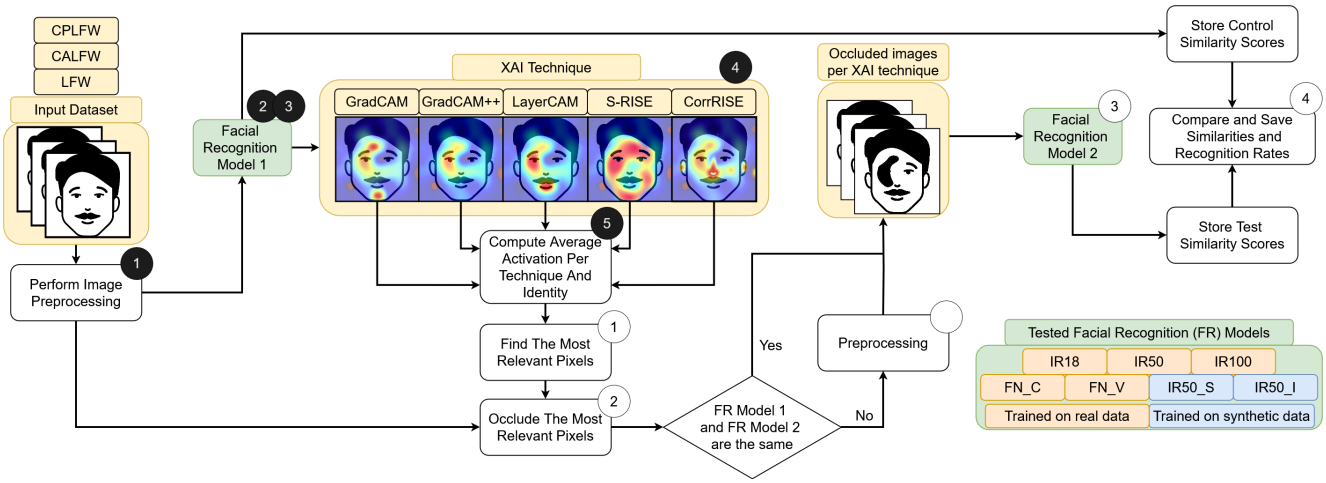

Figure 4. Methodology overview.

RISE [28] uses Monte Carlo sampling to generate random masks and compute importance scores from averaged saliency maps. MC-RISE [16] incorporated the significance of color components into RISE's explanation process, while D-RISE [29] has extended its applicability to object detection. Although effective, perturbation-based methods have a major limitation: they require many occluded image evaluations per sample, which makes them significantly more computationally expensive than CAM-based techniques.

2.2. Adaptations For Facial Recognition

CAM methods cannot be directly applied to facial recognition because the task relies on embedding similarities rather than class scores [25]. Several extensions have therefore been proposed. Grad-CAM adaptations introduced proxy classes or triplet-loss gradients to enable relevance computation in embedding space [2, 11]. More recently, LEAM [6], an adaptation of LayerCAM, used cosine similarity to generate layer-specific explanations. However, its evaluation relied only on decreases in cosine similarity after black-pixel occlusion and did not demonstrate changes in actual recognition performance.

For perturbation-based methods, LIME has been used to highlight visually significant regions [30, 31], although its impact on recognition accuracy was not assessed. S-RISE [24] and CorrRISE [25] adapted RISE to produce saliency maps that distinguish between similar and dissimilar regions for a pair of faces. These studies reuse the insertion and deletion metrics from [28], but offer little insight into which regions truly matter for recognition, or how many pixels must be perturbed to affect the model.

Despite claiming effective region selection, the above work evaluates explanations only at the pair level and does not test whether these regions generalize across images or transfer across models. Consequently, the core properties required for privacy-preserving applications remain unexplored. We address this missing evaluation.

3. Methodology

3.1. Activation Mapping Generation

In the first step of our methodology, we adapt CAM-based techniques to embedding-generating neural networks, which are trained to optimize distances on the hypersphere. Our goal is to obtain activation maps that highlight identity-critical facial regions. We focus on three well-established methods that have already been used for facial recognition explainability [2, 6, 11]: LayerCAM, Grad-CAM, and Grad-CAM++.

For each of these methods, we keep the original activation computation in the forward pass. The main change concerns the gradient computation, which must be adjusted to the embedding setting instead of relying on class scores. While a cosine similarity loss has been used previously [6], we also investigate alternative losses. In particular, we integrate L1 norm, L2 norm, and ArcFace [13] in their positive variants to support mated comparisons.

The activation mapping procedure for CAM-based techniques, illustrated in Figure 4, can be described as follows:

❶ Preprocess (crop, resize, normalize) images.
❷ Obtain embeddings of unperturbed images and activation values via forward pass.
❸ Compute gradients by performing a backward pass with an embedding-oriented loss function.
❹ Compute an activation mapping with a respective CAM-based technique's algorithm.
❺ Repeat the generation for all image pairs and compute the average activation map for each identity.

To complement our experiments, we also re-implement S-RISE and CorrRISE, focusing on positive (similarity-

increasing) regions. This setup enables both within-group and cross-group comparisons between CAM-based and perturbation-based techniques. Due to the unavailability of the original code, both methods were re-implemented directly from the algorithmic descriptions and parameter settings reported in their respective papers. S-RISE uses 1000 random masks, while CorrRISE uses 500 masks and 10 patches of size 30×30 pixels. The full implementation is provided in the replication package.

3.2. Pixel Relevance Evaluation

Evaluating XAI remains challenging: existing methods are sensitive to hyperparameters [3, 41], they often rely on limited or purely qualitative evidence [27, 32, 39], while there is no perfect quantitative metric in the absence of ground-truth relevance maps [17]. Following common practice in classification tasks, we use pixel perturbations, starting from zero-intensity (black) masks [15, 28, 31, 42]. In our setting, this approach is adapted to embedding similarity, and, to reduce dependence on a single perturbation choice, we broaden the occlusion types beyond black pixels by also considering white, random, Gaussian-blurred, and average-value occlusions. Masks are evaluated on the entire dataset, directly assessing the techniques' ability to identify relevant features within the target domain. This design aims to obtain a quantitative and functionally grounded assessment of pixel relevance.

Pixel-relevance evaluation, illustrated in Figure 4, can be summarized as follows:

① Rank pixels using an activation map. Take top x% of them and store their positions.
② Mask detected regions. (One common mask for all images of a given identity)
❍ (Optional) Preprocess (resize, normalize) images for transferability evaluation.
③ Compute embeddings for occluded images.
④ Compare similarity scores and recognition accuracies for occluded and unoccluded images.

3.3. Dataset

We base our experiments on the unconstrained Labeled Faces in the Wild (LFW [18]) dataset alongside specialized cross-age (CALFW [45]) and cross-pose (CPLFW [44]) sets. As one of the standard benchmarks in facial recognition [5], LFW contains 13,233 images of 5,749 individuals and enables direct comparison with prior XAI studies [25, 30]. The number of images per identity varies, reflecting real-world constraints on data availability, and many identities have only a single image. Since our identity-level evaluation requires similarity comparisons across multiple images of the same person, we select 1,680 identities with at least two images to serve as the basis for computing per-identity saliency maps and occlusion masks. Images in CALFW (11,652 images of 4,025 identities) have a larger average age gap (16.61 years), while CPLFW (12,174 images of 3,930 identities) is designed around samples with high pose variability. By analyzing these datasets, we can draw conclusions across diverse conditions and derive more meaningful insights.

3.4. Pretrained Models

We focus on CNN-based models to build on prior work in explainability [27] and to enable direct comparison with existing XAI studies in facial recognition [2, 6, 24, 25], which predominantly employ neural-network architectures. By analyzing well-established CNN architectures and techniques, this study also provides a baseline for future evaluations of vision transformer (ViT) based methods.

We consider seven publicly available pre-trained CNN models (Table 1). These include variants of IResNet [4, 20] and InceptionResnetV1 (FaceNet) [35, 36], To promote research on privacy-friendly alternatives [5], alongside models trained on real facial recognition datasets [1, 9, 40], we also include two models trained entirely on synthetic data: SFace [7] and IDiff-Face [8].

Table 1. Selected models summary.

Abbreviation	Backbone	Training Dataset	Training Data
FN_C	FaceNet	CASIA-WebFace	Real
FN_V	FaceNet	VGGFace2	Real
IR18	IResNet18	Glint360K	Real
IR50	IResNet50	Glint360K	Real
IR100	IResNet100	Glint360K	Real
IR50_S	IResNet50	SFace	Synthetic
IR50_I	IResNet50	IDiff-Face	Synthetic

4. Experiments

We organize our experimental evaluation around the following guiding research questions:

① What proportion of masked pixels is sufficient to induce a meaningful model accuracy reduction?
② Which convolutional layers are the most suitable for relevancy evaluation in gradient-based methods?
③ Which loss functions provide the best performance for the gradient-based methods?
④ Which of the assessed activation mapping generation techniques is the most effective in identifying the most relevant pixels in a black-box scenario?

By assessing the percentage of the image to mask (RQ1), we seek to influence recognition accuracy without sacrificing inconspicuousness in practical contexts. Then, we optimize hyperparameters: layers (RQ2) and loss functions (RQ3) for gradient-based techniques. Due to the compu-

tational complexity of perturbation-based methods, we directly utilize values suggested in the original studies. To conclude the study, we evaluate XAI techniques in both white- and black-box scenarios (RQ4) to assess their generalizability and transferability.

Given the large scale of our evaluation (7 models, 5 occlusion protocols, 3 datasets), the results presented in the main text primarily focus on LFW. We provide additional results and materials in our repository.

4.1. [RQ1] Recognition Performance Against Masked Images

Goal: Due to computational constraints resulting from an extensive evaluation protocol, we estimate the percentage of pixels to be masked to ensure a meaningful performance drop, serving as a baseline for the subsequent RQs.

Experiment: We generate and evaluate activation mappings in a white-box scenario. By changing the threshold of the most relevant pixels to mask – $\{0.5, 1, 2, 3, 5, 10\}$ – we observe the impact on the average cosine similarity and recognition accuracy for the first and last convolutional layer of each model and dataset. The cosine loss function was used throughout this research question.

Results: As presented in Table 2, an increase in the size of occlusion directly translates to a decrease in average cosine similarity and recognition rate for mated images. Nonetheless, the majority of the models is robust to small changes. This outcome can be linked to the findings of [6], which highlight that, beyond a small number of large-intensity outliers, a high quantity of relatively significant pixels is used during facial recognition.

Table 2. LFW recognition accuracy per black pixel occlusion threshold using Grad-CAM on the last convolutional layer. Most significant performance drops (in bold) were observed for FaceNet-based models, with accuracy declining to <0.79 at just 0.5% occlusion.

Model	Threshold [%]					
	0	0.5	1	2	3	5
IR50_I	0.952	0.942	0.942	0.939	0.938	0.935
FN_C	0.949	**0.785**	**0.784**	**0.782**	**0.781**	**0.775**
FN_V	0.965	**0.747**	**0.746**	**0.745**	**0.744**	**0.740**
IR100	0.980	0.975	0.975	0.975	0.975	0.975
IR18	0.976	0.976	0.976	0.976	0.975	0.974
IR50	0.976	0.976	0.976	0.976	0.976	0.976
IR50_S	0.971	0.969	0.969	0.966	0.965	0.960

While the overall impact of occlusions based on the initial layer was minimal, significant drops could be observed for the last convolutional layer when FaceNet-based models were confronted with occlusions generated using Grad-CAM or Grad-CAM++ (Table 3). Perturbation-based techniques – CorrRISE and S-RISE – demonstrated a greater impact against IR50_I trained with synthetic data. While CorrRISE was reported [24, 25] to achieve superior scores of *Insertion* and *Deletion* in a pair-wise task, it does not generalize as well as S-RISE to new samples. This finding demonstrates that techniques that find less impactful, yet still relevant, regions for a specific pair of images can, in fact, generalize better to a wider range of samples and consequently be more useful in domains such as privacy-preserving adversarial attacks.

Table 3. LFW recognition accuracy against black pixel occlusion. The last convolutional layer was used for the gradient-based methods. Most significant performance drops (in bold) were observed mainly for FN_V (0.965→0.540 with Grad-CAM++).

Model	Technique	Threshold [%]			
		0	1	3	5
IR50_I	LayerCAM	0.952	0.952	0.952	0.951
	Grad-CAM	0.952	0.942	0.938	0.935
	Grad-CAM++	0.952	0.951	0.948	0.944
	CorrRISE	0.952	0.931	0.875	**0.791**
	S-RISE	0.952	0.880	**0.814**	**0.791**
FN_V	LayerCAM	0.965	0.965	0.965	0.965
	Grad-CAM	0.965	**0.746**	**0.744**	**0.740**
	Grad-CAM++	0.965	**0.549**	**0.545**	**0.540**
	CorrRISE	0.965	0.955	0.914	0.848
	S-RISE	0.965	0.903	0.839	**0.672**

Further analysis (Table 4) reveals that occlusion choice has a small impact on recognition accuracy, except for blur, which has almost no effect when combined with our approach.

Table 4. Average LFW recognition accuracy per 1% occlusion type. The last convolutional layer was used for the gradient-based methods.

Technique	White	Random	Blur	Mean	Black
LayerCAM	0.966	0.967	0.966	0.966	0.966
Grad-CAM	0.911	0.908	0.966	0.910	0.910
Grad-CAM++	0.767	0.771	0.966	0.768	0.771
CorrRISE	0.966	0.963	0.969	0.966	0.957
S-RISE	0.924	0.930	0.939	0.928	0.923

Having observed a dependence of recognition accuracy on the layer used, in the following research question, we attempt to determine the most optimal setting for tested techniques, using a 1% occlusion threshold. As supported by our findings, this relatively small value can have a noticeable impact on similarity evaluation.

4.2. [RQ2] Layer Importance Comparison

Goal: We aim to uncover which model layer or their combination strategies result in higher values of performance drop among gradient-based techniques.

Experiment: The total number of convolutional layers (denoted as N) is model-dependent – IResNet (21/53/103),

FaceNet (132). We first examine individual layers to determine their impact on the generation of activation maps. We select convolutional layers close to the input (1, 2, 3), at intermediate layers (N/4, N/2, 3N/4) and near the output (N-2, N-1, N). Further, we examine the effects of aggregating multiple layers.

Table 5. LFW cosine similarity per technique (N denotes total number of convolutional layers) against black pixels occlusion. Lowest values (in bold) show no obvious pattern, indicating model-specific behavior.

Model	Technique	Approximate layer position			
		1/4 N	1/2 N	3/4 N	N-1
IR50_I	LayerCAM	**0.575**	0.588	0.586	0.587
	Grad-CAM	0.579	0.586	**0.572**	0.582
	Grad-CAM++	**0.586**	0.591	0.588	0.590
FN_V	LayerCAM	**0.723**	0.744	0.746	0.745
	Grad-CAM	0.727	0.731	0.735	**0.652**
	Grad-CAM++	0.739	0.446	0.745	**0.173**

Table 6. LFW cosine similarity per technique T (LayerCam [L], Grad-CAM [G], Grad-CAM++ [G++]) using averaged activation maps against black pixels occlusion. Similarly to singular layers, the lowest values (in bold) for aggregations show no clear pattern.

Model	T	Layers position(# of layers used)			
		Input(3)	Middle(3)	Output(3)	All(N)
IR50_I	L	**0.579**	0.581	0.589	0.587
	G	0.586	0.586	0.585	**0.584**
	G++	0.588	0.590	**0.588**	0.591
FN_V	L	0.722	0.724	0.730	**0.640**
	G	0.732	0.739	0.713	**0.632**
	G++	0.734	0.745	**0.449**	0.636

Results: Evaluating singular (Table 5) and grouped (Table 6) layers, we have observed that the use of proper layers can yield a much greater model performance drop. We find maximal occlusion susceptibility for FaceNet at the final layers with Grad-CAM and Grad-CAM++ (full results in the supplement). IResNet-based models trained on real data, contrary to their synthetic-based counterparts, struggle exceptionally when the Grad-CAM++ is executed on the second-to-last layer. While the discrepancy for synthetic-data–trained models requires further analysis, SFace was reported [5] to have problems with identity separability and IDiff-Face lacks style variation [26] – two factors that could result in different feature learning dynamics and more distributed model attention, preventing an impact on accuracy. The finding itself has implications for privacy-preserving systems, as it reveals that the use of synthetic data for training could disrupt XAI techniques. Further, tests with averages of all layers and three groups of three – {1, 2, 3}, {N/4, N/2, 3N/4} and {N-2, N-1, N} – combining layers close to the input, in the middle of the network and next to the output showcase that the additional processing cost cannot be justified since only a combination of Grad-CAM++ activations for the last three convolutional layers in FaceNet-based networks resulted in a meaningful performance drop. However, this decline was less severe than the drops observed across most individual layers.

Observed layer relevance was fairly consistent across evaluated occlusion types and close in numerical terms. Overall, layer choice has a negligible influence on models tested against LayerCAM (Table 7), with slightly better results observed for the initial layers. Conversely, Grad-CAM and Grad-CAM++, according to their intended use, are typically the most effective when used with layers close to the output, especially when the second-to-last layer is used. The last layer, the most common choice for these techniques, is a close second. Based on the highest observed performance drops (Table 8), in the following experiments, we will focus only on the best-performing convolutional layers.

Table 7. LFW - top 3 layers ranked by average recognition rate against black pixel occlusion. Best choice in bold.

Technique		Rank		
		1	2	3
LayerCAM	Accuracy	**0.957**	0.958	0.959
	Layer(s)	**3**	1/4 N	Input(3)
Grad-CAM	Accuracy	**0.929**	0.953	0.954
	Layer(s)	**N**	N-2	N-1
Grad-CAM++	Accuracy	**0.639**	0.826	0.863
	Layer(s)	**N-1**	N	1/2 N

Table 8. Best layer per technique T (LayerCam (L), Grad-CAM (G), Grad-CAM++ (G++) across datasets. Black pixel occlusion.

Technique		LFW	CALFW	CPLFW
LayerCAM	Accuracy	0.957	0.836	0.804
	Layer	3	2	2
Grad-CAM	Accuracy	0.929	0.857	0.813
	Layer	N	Input(3)	N-1
Grad-CAM++	Accuracy	0.639	0.525	0.441
	Layer	N-1	N-1	N-1

4.3. [RQ3] Loss Functions Evaluation

Goal: We strive to learn which loss functions, and their respective margins, if applicable, are the most compatible with class activation mapping techniques used in the study.
Experiment: Based on previous findings, we select 1% occlusion threshold and focus on best-performing layers. We evaluate the choice of cosine loss, L1 norm, L2 norm and ArcFace loss functions. Additionally, for ArcFace, we evaluate the impact of the selected margins: {0.0, 0.1, 0.2, 0.3}.

Results: Collected data proves that a technique's ability to find the most relevant pixels is tied to the loss function used (Table 9). Its proper selection can notably reduce recognition accuracy (e.g., substituting cosine loss with 0.2 margin ArcFace for Grad-CAM++: 0.639 $\rightarrow$ 0.575). We refer the reader to the supplementary material for the more detailed results.

Table 9. Average LFW accuracy against black pixels occlusion per loss function with best result per technique T (LayerCam (L), Grad-CAM (G), Grad-CAM++ (G++) in bold.

Loss function	Margin	L	G	G++
L1	-	0.964	0.888	0.594
L2	-	0.964	**0.880**	0.663
Cosine	-	**0.957**	0.929	0.639
ArcFace	0.0	0.963	0.881	0.582
ArcFace	0.1	0.964	0.880	0.595
ArcFace	0.2	0.963	0.880	**0.575**
ArcFace	0.3	0.963	0.880	0.606

Overall, experimental data (Table 10) across the datasets suggest that the LEAM-style [6] cosine loss allows for the greatest, albeit underwhelming, performance drop for LayerCAM. In terms of Grad-CAM++, all tested loss functions provide significant performance drops. Ultimately, ArcFace (margin = 0.2) outperforms other approaches in LFW and CALFW, but underperforms in CPLFW. The closest results have been observed for Grad-CAM, with L2 loss function narrowly surpassing ArcFace in LFW and ArcFace (margin = 0.3) leading in CALFW and CPLFW. Similar to previous findings, occlusion type has a minimal impact on recognition accuracy.

Table 10. Best loss functions for black pixel occlusion per technique T (LayerCam (L), Grad-CAM (G), Grad-CAM++ (G++)) across datasets. Margin for ArcFace is indicated in brackets.

T		LFW	CALFW	CPLFW
L	Accuracy	0.957	0.869	0.807
	Loss func.	Cosine	Cosine	Cosine
G	Accuracy	0.880	0.765	0.665
	Loss func.	L2	ArcFace (0.3)	ArcFace (0.3)
G++	Accuracy	0.575	0.480	0.332
	Loss func.	ArcFace (0.2)	ArcFace (0.2)	L1

Consistent with the previous research question, a much weaker performance drop has been observed across the models trained on synthetic data (IR50_S and IR50_I). The most disruptive occlusions (Table 11) can render the real-data-based models completely useless, while their ethical alternatives remain almost unaffected for LFW and still maintain reasonable accuracies for CALFW and CPLFW.

Table 11. Lowest recognition accuracy among all tested loss function with black pixel occlusions per model and dataset. Results > 0.5 in bold.

Dataset	IR50_S	IR50_I	FN_C	FN_V	IR100	IR18	IR50
LFW	**0.965**	**0.925**	0.419	0.079	0.425	0.448	0.450
CALFW	**0.831**	**0.830**	0.451	0.055	0.277	0.356	0.299
CPLFW	**0.684**	**0.669**	0.211	0.058	0.111	0.217	0.144

4.4. [RQ4] Transferability of Activation Mapping Techniques

Goal: Building on RQ1–3, we compare saliency methods to determine which approach to per-identity pixel selection achieves the highest generalization and transferability.
Experiment: Optimal layers (RQ2) and loss functions (RQ3) for gradient-based techniques were chosen as indicated in previous research questions, while parameter values from original studies were used for perturbation-based techniques. The dataset size is reduced to 500 random identities for perturbation-based techniques due to computational complexity. Using 3 datasets, 4 occlusion types (blur was discarded due to minimal impact in previous RQs), 7 models for key regions selection and then evaluation, each technique is tested in 504 black-box (distinct models used for occlusion generation and evaluation) and 84 white-box (the same model used for both tasks) configurations.
Results: Analyzing outputs per technique, we have observed quite consistent results with respect to methods' overall efficacy. The largest magnitudes of recognition performance drop have been observed for Grad-CAM++ (Table 12), followed by the limited impact of Grad-CAM. LayerCAM, S-RISE, and CorrRISE have an overall marginal impact on recognition performance, suggesting that the regions highlighted by those techniques are far from being the most relevant ones.

Table 12. Average performance per technique across datasets using black pixel occlusions

Technique	LFW	CPLFW	CALFW
GradCAM++	0.622	0.375	0.485
GradCAM	0.851	0.661	0.748
S-RISE	0.945	0.797	0.853
CorrRISE	0.954	0.817	0.859
LayerCAM	0.958	0.812	0.847

To confirm the statistical significance of our findings, we have performed Cohen's effect size measure and paired t-test (occluded vs. unoccluded). As presented in Table 13, our results are statistically significant ($p < 0.001$). For the LFW dataset, which exhibits the smallest performance drop, the effect sizes vary significantly. We observe a large effect for Grad-CAM++ (d = 1.387), a medium effect for Grad-CAM (d = 0.492), and small effects for the others ($d < 0.2$).

Table 13. Per-identity statistical evaluation of 1% black pixel occlusion recognition based on LFW accuracy results. All techniques show statistically significant differences ($p < 0.001$) with effect sizes ranging from small to large (Cohen's d: 0.096-1.387).

Technique	Accuracy	p-value	Cohen d
Grad-CAM++	0.622	<0.001	1.387
Grad-CAM	0.851	<0.001	0.492
S-RISE	0.945	<0.001	0.130
CorrRISE	0.954	<0.001	0.134
LayerCAM	0.958	<0.001	0.096

Table 14. Activation mapping transferability of Grad-CAM++ against black pixels occlusion in CPLFW. Lower values indicate stronger transferability. Values below 0.5 are in bold.

Generation model	Evaluation model						
	IR50_I	FN_C	FN_V	IR100	IR18	IR50	IR50_S
IR50_I	0.759	0.898	0.936	0.935	0.935	0.765	0.852
FN_C	**0.284**	**0.337**	**0.332**	**0.392**	**0.351**	**0.284**	**0.326**
FN_V	**0.042**	**0.075**	**0.051**	**0.147**	**0.081**	**0.052**	**0.077**
IR100	**0.112**	**0.161**	**0.146**	**0.232**	**0.173**	**0.120**	**0.158**
IR18	**0.142**	**0.195**	**0.181**	**0.261**	**0.209**	**0.150**	**0.189**
IR50	**0.122**	**0.176**	**0.162**	**0.244**	**0.188**	**0.132**	**0.170**
IR50_S	0.761	0.898	0.936	0.934	0.934	0.766	0.852

Analyzing the data presented in Table 14 and the supplementary material, we have uncovered multiple interesting pattens: (1) Transferability – black-box attacks can be more successful than white-box. For exceptionally successful patterns, the accuracy scores are directly linked to the model used to generate activation rather than the model to which the attack was transferred; (2) Models trained on synthetic data (IR50_I and IR50_S), which proved remarkably resistant in white-box attacks, are not resistant to transferred occlusions. Their behavior is consistent in scale with other IResNet models trained on real data; (3) IResNet models, trained on real data, when used for occlusion generation, provide a similar magnitude of recognition accuracy regardless of their size. FaceNet-based models trained on distinct datasets tend to differ significantly in performance.

Table 15. Average LFW accuracy across occlusion types.

Technique	Black	White	Mean	Random
GradCAM++	0.622	0.622	0.585	0.585
GradCAM	0.851	0.851	0.839	0.839
S-RISE	0.945	0.922	0.942	0.929
CorrRISE	0.954	0.962	0.961	0.961
LayerCAM	0.958	0.958	0.957	0.956

Finally, occlusion type selection has a minor impact on recognition accuracy (Table 15). Black and white pixels usually yield a smaller performance drop, suggesting model robustness to extreme values.

5. Discussion

Our results enable a clearer interpretation of several claims made in prior work. Chen et al. [11] previously observed that Grad-CAM++ adaptations yield more accurate relevance maps than Grad-CAM when applied to embedding networks. Our findings confirm this observation (Table 13) across a broader set of models and occlusion protocols, indicating that higher-order gradient weighting remains consistently more reliable in facial recognition settings.

Borsukiewicz et al. [6] introduced LEAM as a LayerCAM-based adaptation using a cosine similarity loss, with an emphasis on early convolutional layers. While our experiments confirm proper selection of target layers (Table 7) and loss function (Table 9), LEAM's measured influence on recognition accuracy is negligible, as shown by Cohen's d=0.096 (Table 13), raising questions about the practical utility of LEAM and suggesting that cosine-similarity drops alone are insufficient indicators of region relevance.

For perturbation-based approaches, S-RISE and CorrRISE [24, 25] were proposed to analyze similar and dissimilar regions between face pairs. CorrRISE achieves strong insertion and deletion scores. However, our evaluation shows that its explanations do not generalize across images when a fixed occlusion set must be reused. S-RISE shows slightly better adaptability but remains substantially less aligned with recognition behaviour than Grad-CAM and Grad-CAM++ (Table 13).

Overall, these comparisons reveal a common pattern: methods designed around pairwise explanations cannot capture the identity-level consistency required for generalizable and transferable explanations, whereas Grad-CAM++ is the only method whose highlighted regions remain predictive of recognition performance across identities and models.

6. Conclusion

We evaluated identity-level generalizability and cross-model transferability of facial-recognition explanations across multiple methods, models, and occlusion settings. Identified hyperparameters strengthened identity-level effects and showed that key regions transfer well across models, especially important for those trained on synthetic data, where XAI methods are otherwise less effective. These findings lay the groundwork for XAI techniques to guide adversarial attacks in the digital domain, helping preserve individual privacy online. Future work will extend this evaluation to transformer-based methods and explore the impact of pose variability across images.

Acknowledgments

- This research was funded by the Luxembourg Army.
- The experiments were primarily carried out using the HPC facilities of the University of Luxembourg.

References

[1] Xiang An, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine, 2021.

[2] Mudit Bachhawat. Generalizing gradcam for embedding networks, 2024.

[3] Naman Bansal, Chirag Agarwal, and Anh Nguyen. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 8673–8683, 2020.

[4] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International conference on machine learning*, pages 573–582. PMLR, 2019.

[5] Paweł Borsukiewicz, Fadi Boutros, Iyiola E. Olatunji, Charles Beumier, Wendkûuni C. Ouedraogo, Jacques Klein, and Tegawendé F. Bissyandé. Beyond real faces: Synthetic datasets can achieve reliable recognition performance without privacy compromise, 2025.

[6] Paweł Jakub Borsukiewicz, Jordan Samhi, Jacques Klein, and Tegawendé F Bissyandé. Explainable ai for analyzing person-specific patterns in facial recognition tasks. *arXiv preprint arXiv:2509.17457*, 2025.

[7] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022.

[8] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023.

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[10] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017.

[11] Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting grad-cam for embedding networks. *CoRR*, abs/2001.06538, 2020.

[12] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[14] European Union. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act), 2024.

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

[16] Yuhki Hatakeyama, Hiroki Sakuma, Yoshinori Konishi, and Kohei Suenaga. Visualizing color-wise saliency of black-box image classification models. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[17] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M-C Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[18] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[19] Rami Ibrahim and M Omair Shafiq. Explainable convolutional neural networks: a taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):1–37, 2023.

[20] InsightFace. Insightface model zoo, 2021. Accessed: 2025-09-27.

[21] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps. *IEEE Transactions on Image Processing*, PP:1–1, 2021.

[22] Martin Knoche, Torben Teepe, Stefan Hörmann, and Gerhard Rigoll. Explainable model-agnostic similarity and confidence in face verification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 711–718, 2023.

[23] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1046–1055, 2021.

[24] Yuhang Lu and Touradj Ebrahimi. Explanation of face recognition via saliency maps. In *Applications of Digital Image Processing XLVI*, pages 218–229. SPIE, 2023.

[25] Yuhang Lu, Zewei Xu, and Touradj Ebrahimi. Towards visual saliency explanations of face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4726–4735, 2024.

[26] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Qiuyang Yuan, Xuan Zhao, Jianqing Xu, Shouhong Ding, Shaoming Wang, Rizen Guo, and Shuigeng Zhou. Data synthesis with diverse styles for face recognition via 3dmm-guided diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21203–21214, 2025.

[27] 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*, 55(13s): 1–42, 2023.

[28] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[29] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11443–11452, 2021.

[30] Ankit Rajpal, Khushwant Sehra, Rashika Bagri, and Pooja Sikka. Xai-fr: explainable ai-based face recognition using deep neural networks. *Wireless Personal Communications*, 129(1):663–680, 2023.

[31] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[32] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pages 45–50, 2021.

[33] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[34] Sédrick Stassin, Alexandre Englebert, Géraldin Nanfack, Julien Albert, Nassim Versbraegen, Gilles Peiffer, Miriam Doh, Nicolas Riche, Benoît Frenay, and Christophe De Vleeschouwer. An experimental investigation into the evaluation of explainability methods for computer vision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 91–106. Springer, 2023.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[36] Timesler. facenet-pytorch, 2018. Accessed: 2025-10-01.

[37] Giorgio Visani, Enrico Bagli, and Federico Chesani. Optilime: Optimized lime explanations for diagnostic computer algorithms. *arXiv preprint arXiv:2006.05714*, 2020.

[38] Lynn Vonder Haar, Timothy Elvira, and Omar Ochoa. An analysis of explainability methods for convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 117:105606, 2023.

[39] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[41] Gal Yona and Daniel Greenfeld. Revisiting sanity checks for saliency maps. *arXiv preprint arXiv:2110.14297*, 2021.

[42] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[43] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. In *Uncertainty in artificial intelligence*, pages 887–896. PMLR, 2021.

[44] Tianyue Zheng and Weihong Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7):5, 2018.

[45] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[46] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[47] Zhengze Zhou, Giles Hooker, and Fei Wang. S-lime: Stabilized-lime for model explanation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2429–2438, 2021.