Towards Visual Saliency Explanations of Face Verification

Yuhang Lu, Zewei Xu, and Touradj Ebrahimi
EPFL, Lausanne, Switzerland
firstname.lastname@epfl.ch

Abstract

In the past years, deep convolutional neural networks have been pushing the frontier of face recognition (FR) techniques in both verification and identification scenarios. Despite the high accuracy, they are often criticized for lacking explainability. There has been an increasing demand for understanding the decision-making process of deep face recognition systems. Recent studies have investigated the usage of visual saliency maps as an explanation, but they often lack a discussion and analysis in the context of face recognition. This paper concentrates on explainable face verification tasks and conceives a new explanation framework. Firstly, a definition of the saliency-based explanation method is provided, which focuses on the decisions made by the deep FR model. Secondly, a new model-agnostic explanation method named CorrRISE is proposed to produce saliency maps, which reveal both the similar and dissimilar regions of any given pair of face images. Then, an evaluation methodology is designed to measure the performance of general visual saliency explanation methods in face verification. Finally, substantial visual and quantitative results have shown that the proposed CorrRISE method demonstrates promising results in comparison with other state-of-the-art explainable face verification approaches.

1. Introduction

Recent years have witnessed great advances in face recognition (FR) due to the rapid development of deep learning techniques. Current deep face recognition systems achieve near-perfect performance on well-known public benchmarks and have been widely deployed in several applications, such as access control and surveillance. However, the predictions made by deep learning-based systems tend to be challenging to interpret. The deployment of such biometric systems poses a potential threat to privacy and data protection, resulting in serious public concern. To address these issues, it is essential to comprehend the decision-making process of deep face recognition, thereby improving their performance and making them more widely accepted in society.

This paper focuses on a crucial problem in face recognition, i.e. explainable face verification, specifically by developing insightful explainability tools to interpret the decision-making process of a deep face verification system. Various saliency map-based algorithms have been proposed as forms of explainable artificial intelligence (XAI) to highlight either the internal CNN layers [2, 23, 35] or the important pixels of the input image that are relevant to the model's decision [3, 5, 18, 27, 29, 36, 38]. While many of them have achieved impressive results, they are mainly designed for classification and detection tasks.

Face verification differs from other vision tasks not only due to the notable difference in the output format, but also the decision-making process, which often involves two images. Explaining an FR model does not mean simply highlighting the critical areas using importance maps but, beyond this, it should also interpret why the given face images are matching or non-matching to the verification system [31]. In this context, we first set out and improve the mainstream definition of visual saliency map-based explanations for face verification systems. Specifically, an explanation method should reveal the similar regions when the FR system believes the input images are matching and the dissimilar regions if they are non-matching. Then a Correlation-based Randomized Input Sampling for Explanation (CorrRISE) algorithm is proposed, which is model-agnostic and capable of highlighting both the similar and dissimilar regions between any two input face images. In addition, the paper proposes a new objective evaluation methodology to compare different state-of-the-art explainable face recognition (XFR) methods in a quantitative manner. The contributions of this paper can be summarized as follows:

- A model-agnostic explanation method called CorrRISE is proposed to highlight the similarity and dissimilarity regions between any two face images.
- A new evaluation methodology is conceived to quantitatively measure the performance of general saliency map-based explanation methods for face verification.

Support from XAIface CHIST-ERA-19-XAI-011 and the Swiss National Science Foundation (SNSF) 20CH21_195532 is acknowledged.

- Extensive experiments on multiple face verification scenarios have been carried out and presented, demonstrating the effectiveness of the proposed method.

2. Related Work

Visual saliency algorithms have been widely used to explain decision systems in vision tasks that rely on deep learning techniques. A saliency map is essentially an image where each pixel value represents the importance of the corresponding pixel. This map helps identify the significant areas of an input image that contribute to the final output of a "black-box" model. In general, there are two types of methods to create such saliency maps.

The first category of methods involves backpropagating an importance score from the output of the model to the input pixels through the neural network layers. Earlier work includes Gradient Backpropagation [29], Layer-wise Relevance Propagation [3], Class Activation Maps (CAM) [38], and Excitation Backpropagation [36]. This type of approach often requires access to the intrinsic architecture or gradient information of the deep model. Grad-CAM [28] and Grad-CAM++ [5] generalize CAM to be applied to arbitrary convolutional neural networks (CNNs) by weighing the feature activation values with the class-specific gradient information that flows into the final convolution layer. Considering their fame and flexibility in CNNs-based models, this work adapts them to the explainable face verification problem to compare with our proposed method.

The second category of methods performs random perturbations on the input image, such as adding noise or occlusion, and produces saliency maps by observing the impact on the model's output [6, 10, 24, 26]. For example, Ribeiro et al. [26] proposed an interpretable approximate linear decision model (LIME) in the vicinity of a particular input which analyzes the relation between the input data and the prediction in a perturbation-based forward propagation. The RISE [24] algorithm generates random binary masks, applies them to the input image, and then uses the output class probabilities as weights to compute a weighted sum of the masks as a saliency map. Our method employs a similar random-masking idea, but differently from RISE, we address a more challenging problem of explainable face verification and propose to incorporate correlation operations into the saliency maps generation process.

While most XAI techniques involving saliency are developed for image classification, there is a growing demand for explanation methods in other image understanding tasks, including object detection [25] and image similarity search and retrieval [8, 13, 30]. In face recognition, earlier work [4, 31, 32] mainly adapted saliency-based explanation algorithms [24, 27, 27, 36] from classification tasks. Alternative research work focuses on face verification models that are explainable by themselves, often referred to as intrinsic explanation methods. For example, Yin et al. [34] designed a feature activation diverse loss to encourage learning more interpretable face representations. Lin et al. [19] proposed a learnable module that can be integrated into face recognition models and generate meaningful attention maps. But these methods need to be trained exclusively and thus are not feasible for those deployed verification systems by third parties. Instead, more recent studies [21,22] provide "black-box" explanations to arbitrary face verification models. The authors introduced several perturbation-based methods to create explainable saliency maps without manipulating or retraining the model and achieved visually promising results. xFace [17] further improved it by applying more systematic occlusions to inputs and measuring the feature distance deviations. However, the evaluation of all the above-mentioned approaches has been based on visualizations, making it difficult to compare to other. This paper provides an objective evaluation methodology and compares our explanation approach with the above-mentioned state-of-the-art model-agnostic methods in a quantitative manner.

3. Proposed Method

3.1. Definition of Explanation for Face Verification

In face verification, a deep face-matching model operates by comparing a predefined threshold value with the similarity score between two face images. Ideally, an explanation method should give a visual interpretation of the deep model's decision.

An earlier study [4] defined explainability in face verification tasks as a way to visualize the discriminative information of every single face. But practically, the decision-making process of the face recognition model often involves two or more images. The regions that are similar to the FR model may not naturally be the global critical areas. Williford et al. [31] have leveraged a face triplet, i.e. probe, mate, and nonmate, to explain the relative importance of facial regions. They defined explainable face matching as a way to highlight the regions of the probe image that maximize the similarity with the mate image and minimize the similarity with the nonmate. However, face recognition systems make decisions independently for each pair of inputs instead of a triplet. [21,22] improved the definition by directly exploring the relevant parts between two images when the matching is established. But the irrelevant parts between the inputs are not considered, which dominates the decision-making process for non-matching pairs.

This paper summarizes the problem of explainable face verification as follows. Given a pair of images feeding into the deep face-matching model, the explanation method should produce the corresponding saliency maps for both input images respectively, which help clearly interpret the prediction results by answering the following questions:

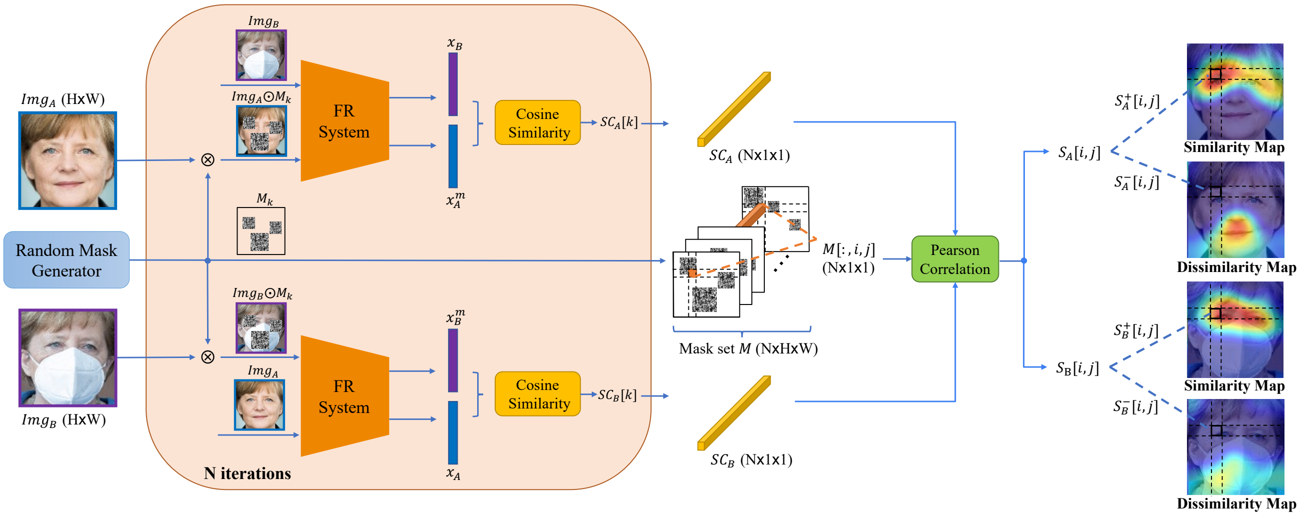

Figure 1. Workflow of the proposed CorrRISE explanation method. The similarity and dissimilarity maps are calculated respectively given an arbitrary input face pair. The block in the middle repeats N iterations using different random masks. The output similarity scores and the mask set are fed to the correlation module to calculate similarity and dissimilarity saliency maps in a pixel-wise manner.

- If the face verification system believes the input pair is matching, which regions are **similar** to the model?
- If the face verification system believes the input pair is non-matching, which regions are **dissimilar** to the model?

3.2. Correlation-based Randomized Input Sampling for Explanation (CorrRISE)

Most existing explanation methods for face verification interpret the model's prediction with saliency maps that indicate similar regions between any matching images, but they are not eligible for generating meaningful saliency maps for non-matching cases. This section presents a new model-agnostic explanation method called CorrRISE to address the new definition of explainable face verification, see Fig. 1. In principle, CorrRISE generates saliency maps by injecting perturbation and observing the impact on output. Thus, it provides "black-box" explanations and can be applied to any FR system without retraining or access to the network. In contrast with other perturbation-based approaches explaining classification models, CorrRISE applies random masks to face images and measures the effect of masked regions on the final similarity scores between two faces rather than a single categorical output. Then, the Pearson correlation between a list of similarity scores and random masks is calculated in a pixel-wise manner to obtain saliency maps. The similar and dissimilar pixels are disentangled from the saliency map according to the correlation coefficients. This is the key innovation of CorrRISE and distinguishes it from previous explainable face verification approaches and prior RISE adaptations by [21, 31]. In detail, the CorrRISE algorithm comprises two pivot steps, i.e. mask generation and saliency map generation.

3.2.1 Mask Generation

The original RISE algorithm samples binary masks and upsamples them to larger resolutions with bilinear interpolation so that they have values from 0 to 1. Mery [21] adapted RISE to face verification tasks and generated masks following Gaussian distribution. We simplify this process here by randomly generating multiple small square patches in various locations of a plain image. The values of each patch are not linearly distributed but randomly sampled between $[0, 1]$. The parameters are reported in the experimental section. Basically, the mask generation steps are as follows:

1. Initialize the parameters of the mask generator, i.e. the total number of masks N, and the number and size of square patches in each mask.
2. Sample multiple square patches with arbitrary values between $[0, 1]$ in random locations of each mask M_i and finally get the mask set $\{M_i, i = 1, ..., N\}$.

3.2.2 Correlation-based Saliency Map Generation

Fig. 1 illustrates an overview of the proposed CorrRISE method and Algorithm 1 presents detailed steps. In general, given a pair of images $\{Img_A, Img_B\}$ and a face recognition model f_x, the objective is to produce saliency maps that highlight both the similar and dissimilar regions between the two faces. First, CorrRISE leverages a mask generator

Algorithm 1

```
procedure CorrRISE
    Input: Number of iterations $N$, FR model $f_x$, Similarity function Score(), Face images $I_A$ and $I_B$
    Output: Saliency maps $S_A^+, S_A^-, S_B^+, S_B^-$
    $H, W \leftarrow$ Size($I_A$)
    for $k = 1 : N$ do
        $M_k \leftarrow$ RandomMaskGenerator($H, W$)
        $x_A, x_B \leftarrow f_x(I_A), f_x(I_B)$
        $x_A^m, x_B^m \leftarrow f_x(I_A \odot M_k), f_x(I_B \odot M_k)$
        $SC_A[k] \leftarrow$ Score($x_A^m, x_B$)
        $SC_B[k] \leftarrow$ Score($x_B^m, x_A$)
        $M[k, :, :] \leftarrow M_k$
    end for
    for $i = 1 : H$ do
        for $j = 1 : W$ do
            $S_A[i, j] \leftarrow$ PearsonCorr($SC_A, M[:, i, j]$)
            $S_B[i, j] \leftarrow$ PearsonCorr($SC_B, M[:, i, j]$)
        end for
    end for
    $S_A^+, S_A^- \leftarrow S_A[S_A \geq 0], S_A[S_A < 0]$
    $S_B^+, S_B^- \leftarrow S_B[S_B \geq 0], S_B[S_B < 0]$
end procedure
```

to randomly produce N masks $M = \{M_i, i = 1, ..., N\}$. Each of the masks M_i will be multiplied with the input images respectively, e.g. img_A. The masked $img_A \odot M_i$ and unmasked img_B are fed into the face recognition model f_x to capture the deep face representation $\{x_A^m, x_B\}$. The cosine similarity score SC_i between the deep features is then calculated. After iterating all the N masks, the list of scores $SC = \{SC_i, i = 1, ..., N\}$ corresponding to the mask list is recorded. Subsequently, Pearson correlation is performed between SC and M in a pixel-wise manner to obtain the final saliency map S_A for img_A. The location of positive correlation coefficients represents the regions on img_A that are similar to img_B, while the location of negative coefficients represents the dissimilar regions. The same procedure is applied to img_B to get the saliency map S_B. We perform these saliency map generation procedures separately for two images, see Fig. 1, because a face-matching system will verify two irrelevant but masked faces as a matching pair, which interferes generation process.

4. Evaluation Methodology

Despite the promising development of explainability methods in vision tasks, the importance of rigorous objective evaluation methodologies has long been overlooked. In particular, only a few metrics have been designed for visual saliency explanation tools. In the past, human evaluation was the predominant way to assess model explainability [12, 36]. Petsiuk et al. [24] proposed two automatic objective metrics, "Deletion" and "Insertion", for explainable image classification task, which measure the change in output classification probability after removing or adding salient pixels from/to the input image. [13] adapted these metrics to image retrieval tasks. In face verification, [4] quantified the visualized discriminative features by playing a "hiding game", which iteratively obscures the least important pixels in the image sorted according to a produced attention map. But it is not precise enough to differentiate high-performing explanation methods [33].

This paper proposes new "Deletion" and "Insertion" metrics to better fit explanation methods for face verification. Specifically, these metrics measure the change in the verification accuracy after modifying the input image according to the importance map generated by the explanation method. The intuition behind the proposed metrics is that the explanation saliency map is expected to precisely highlight the most important regions of two faces with the smallest number of pixels, based on which the FR model can make final decisions. The faster the overall accuracy drops/rises after removing/adding salient pixels, the more accurate the produced saliency map.

Algorithm 2 Deletion and Insertion Metric

```
procedure Evaluation Metric
    Input: FR model $f_x$, FR evaluator eval(), testing dataset $D$, saliency maps $S$, number of steps $n$
    Output: deletion score $d_1$, insertion score $d_2$
    $step \leftarrow 1/n \times 100$
    $S \leftarrow$ Sorting($S$)                ▷ sort in descending order
    for $i = 1 : n$ do
        $p \leftarrow i/n \times 100$
        Mask the first $p\%$ pixels of $D$ to get $D_1'$
        Insert the first $p\%$ pixels to plain images to get $D_2'$
        $acc_{1i} \leftarrow$ eval($f_x, D_1'$)
        $acc_{2i} \leftarrow$ eval($f_x, D_2'$)
    end for
    $d_1 \leftarrow$ AreaUnderCurve($acc_{1i}$ vs. $i/n,\ i = 1 : n$)
    $d_2 \leftarrow$ AreaUnderCurve($acc_{2i}$ vs. $i/n,\ i = 1 : n$)
end procedure
```

Algorithm 2 describes the procedure for calculating the proposed metrics. In general, the "Deletion" metric executes the following steps. First, the generated saliency map for each input image is sorted according to the importance value. Then, n threshold values are evenly sampled from $[0, 1]$, i.e. $p_k = k/n \times 100\%$ where $k \in \{1, \ldots, n\}$, and serve as different percentages of modified pixels in the input image. Subsequently, a verification task using a dataset D is performed iteratively as an auxiliary task in the overall evaluation methodology. In each iteration, p_k percent of the most salient pixels are masked from the input image of the entire testing dataset. The accuracy of the face recognition model is then measured on the modified dataset. Finally, the "Deletion" metric is defined as the AUC score of the Accuracy vs. Percentage of masked pixels curve. The lower the metric, the more accurate the evaluated saliency map. The "Insertion" metric takes a complementary approach but performs similar procedures. In each iteration, p_k percent of the most important pixels are inserted into a plain image. Ideally, a higher "Insertion" score refers to a better explanation of the saliency map.

5. Experimental Results

5.1. Implementation Details

5.1.1 Explanation Method Setup

The proposed CorrRISE explanation method does not require any training or access to the internal architecture of the face recognition model. During the explanation process, the number of generated masks, i.e. number of iterations, is set to 500 by default. For each mask, there are 10 patches and the size of each patch is 30×30 pixels.

As a comparison, two state-of-the-art XFR methods MinPlus [21], and xFace [17] have been tested based on their official open-source code. Meanwhile, several XAI methods [5, 24, 26, 27] have been adapted and tested. For Grad-CAM [27] and Grad-CAM++ [5], instead of back-propagating the gradients of class-wise posterior probability to activation layers, we adapt them by performing back-propagation for gradients of similarity scores between two input images. Moreover, we utilize the third-party adaptation from authors of [21] for LIME [26] and RISE [24].

5.1.2 Face Recognition Model Setup

This paper adopts three different state-of-the-art face recognition models for experiments to show the validity of our proposed model-agnostic method, namely ArcFace [7], MagFace [20], and AdaFace [16]. All the models share the same ResNet-50 [11] feature extractor and are trained on the same dataset by running the publicly available code. Except for the cross-model analysis, all the experiments are conducted with the ArcFace model.

5.1.3 Dataset

The face recognition models are trained with the MS1Mv2 dataset [7]. For evaluation, this paper first selects samples from different datasets, namely LFW [15], CPLFW [37], LFR [9], and Webface-Occ [14] for visual presentation in various verification scenarios. In the proposed objective evaluation methodology, LFW, CPLFW, and Webface-Occ datasets are used for quantitative evaluation. All the images are cropped and resized to 112x112 pixels.

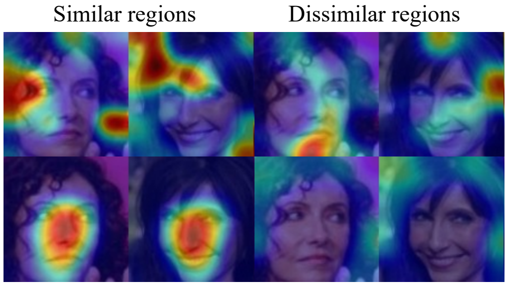

Figure 2. Sanity check for the CorrRISE explanation method. The first row is the explanation heatmap for a deep model with randomized parameters, while the second row is for a normal face recognition model. The importance increases from blue to red color.

5.2. Sanity Check

A recent study [1] has raised doubts about the reliability of visual saliency methods that the produced explanation heatmaps can be independent of the deep model or the input data. They introduced a model parameter randomization test for a sanity check. In the context of face verification, an explanation method may provide visually compelling heatmaps by directly emphasizing the center of the faces without interacting with the deep FR model. Therefore, this paper employs a similar sanity check to validate the effectiveness of the proposed method. Specifically, the parameters of the ResNet-50 backbone network are randomly initialized and then the CorrRISE algorithm is applied to the randomized model. The first row of Fig. 2 shows that the CorrRISE algorithm will generate nonsense saliency maps when it tries to explain a face recognition model with fake parameters. It proves that the proposed explanation method relies on a feasible recognition model and is capable of producing meaningful interpretations.

5.3. Visual Explanation Results

This section presents the visual results of the saliency maps generated by our proposed CorrRISE algorithm. For easier demonstration and fair comparison, all the experiments here are conducted on the ArcFace model. First, the explanation ability of CorrRISE is tested on the standard face verification scenario with samples randomly selected from LFW and WebFace-Occ datasets respectively. Then, the behavior of the ArcFace model in several challenging scenarios is analyzed and explained, notably some failure cases due to very similar subjects or significant head pose variations. Moreover, a visual comparison with other explanation approaches is presented.

5.3.1 Standard Verification Scenario

In face verification, a deep FR model predicts whether a pair of face images belong to the same identity. Fig. 3 shows the

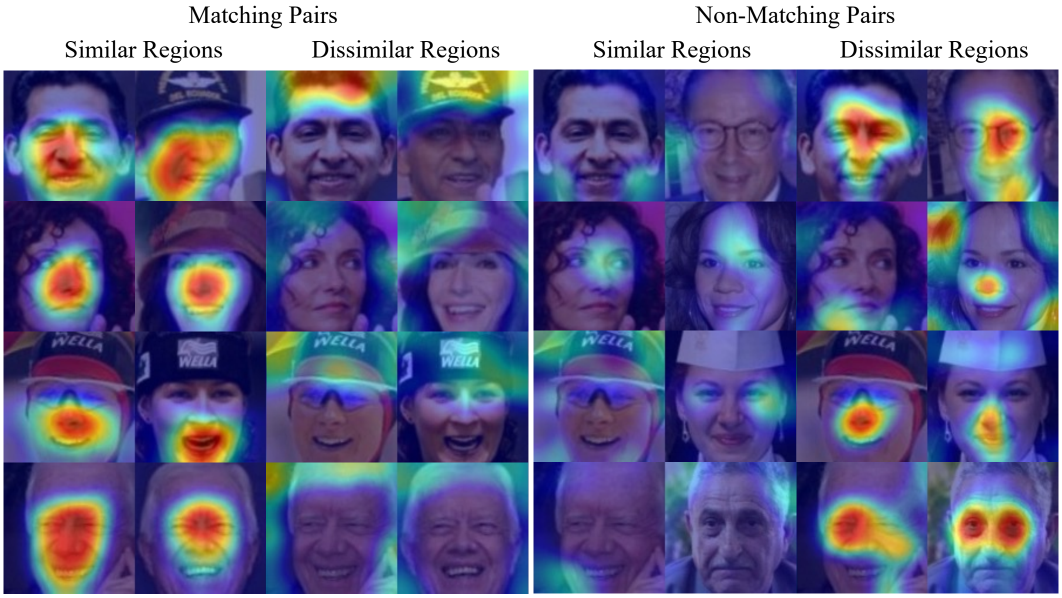

Figure 3. Visual explanation results from CorrRISE for both matching and non-matching face pairs in standard face verification scenario. The produced saliency maps explain why the verification model makes correct predictions on all face pairs. The saliency value increases from blue to red color.

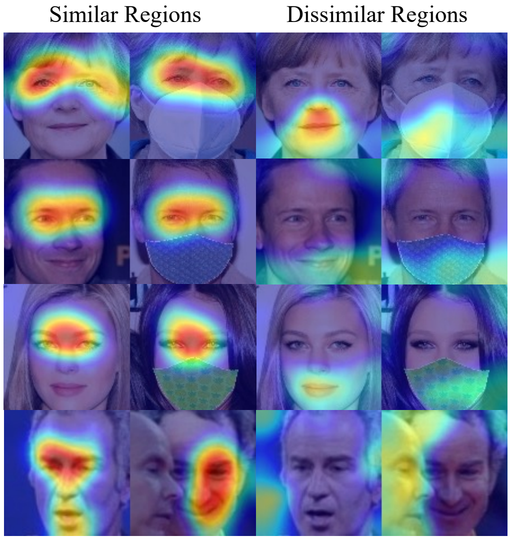

Figure 4. Saliency map explanations for the predictions of the FR model on partially-occluded faces. The masked regions are accurately identified as dissimilar regions. The saliency value increases from blue to red color.

visual explanations for the model's decision on four matching and four non-matching pairs of images taken from the LFW dataset. Notably, here the deep model makes correct predictions on all eight pairs.

As a result, the saliency maps produced by CorrRISE properly highlight regions that the FR model believes are very similar between the matching pairs. The general salient region focuses on eyes, noses, and mouth, but it also varies from person to person. For example, the FR model relies more on the cheek when comparing the matching pair in the first row while more on the open mouth for the pair in the third row. It is also notable that the dissimilar regions often concentrate on irrelevant backgrounds and unexpected occlusions, such as hats. As for the non-matching pairs, CorrRISE tries to localize the similar areas between the non-matching faces but with rather low salient values. In contrast, it produces saliency maps that clearly spotlight the most dissimilar regions in their faces, indicating very low similarity between them.

To further show the effectiveness of the proposed explanation algorithm, an additional test has been performed with partially occluded faces, taken from WebFace-Occ. In this experiment, the face recognition model is also capable of verifying occluded faces despite slightly lower similarity scores. As presented in Fig. 4, the CorrRISE algorithm precisely localizes the non-occluded regions which the FR model leverages for correct predictions, whilst it also highlights the occlusions, such as masks and even another person's face, as dissimilar regions.

5.3.2 Challenging Scenarios

Face verification systems can suffer from various challenging situations in daily usage, such as very similar identities or head pose changes. It is important to provide reliable explanations for the system's behavior in specific scenarios.

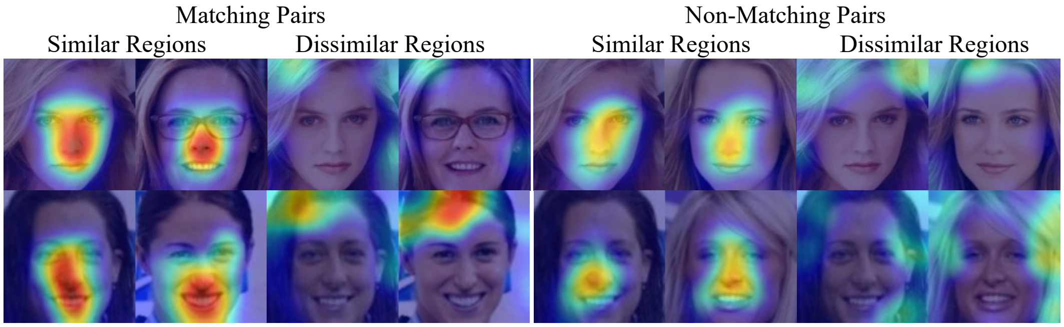

Figure 5. Saliency map explanation that interprets why the FR model mistakenly matches the two non-matching pairs (right). The saliency value increases from blue to red color.

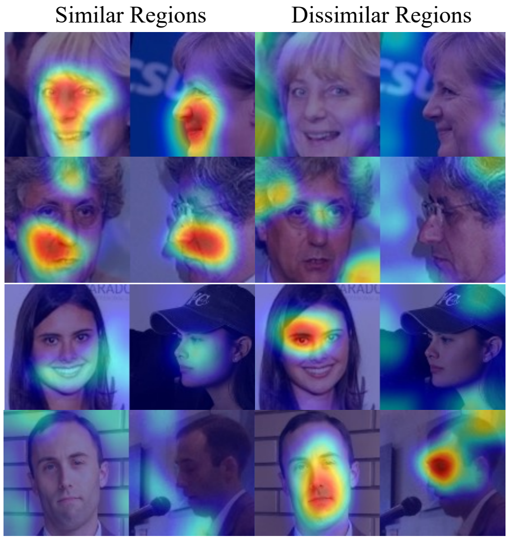

Figure 6. Saliency map explanations for the predictions of the FR model on matching but ill-posed faces. The FR model makes correct predictions in the first two examples and fails in the rest. The saliency value increases from blue to red color.

Fig. 5 presents two examples of similar identity scenarios in a format of triplets. The deep model has mistakenly verified both the matching (left) and non-matching pairs (right) as faces from the same subject. This wrong decision is explained by the saliency maps produced by CorrRISE. They show that, although with lower salient values than the matching pairs, the deep model believes the nose and mouth regions of the two non-matching images are close enough to be classified as the same person. Meanwhile, there is no significant dissimilar region between them.

Head pose variation is another well-known challenge for face verification systems. Fig. 6 shows four examples, where the deep model correctly recognizes the first two while failing at the last two. The saliency maps corresponding to the first two examples show that the deep model manages to localize similar regions with high salient values even on the sided faces. In contrast, it fails to find enough similarities in the last two examples and makes false predictions due to lacking sufficient information. For instance, the dissimilar saliency map spotlights the left eye of the front face in the third example, which corresponds to the missing parts in the sided face.

5.3.3 Comparison with other Explanation Approaches

A visual comparison is made between the proposed method and six explainable face verification approaches, including adaptations of four XAI approaches, i.e. Grad-CAM [27], Grad-CAM++ [5], LIME [26], and RISE [24], and two state-of-the-art XFR methods, namely MinPlus [21] and xFace [17]. Both MinPlus and xFace consist of several variations and the best-performing one is selected. Implementation details refer to Section 5.1.1. Due to the limitation of several methods, only the similar regions between two images are visualized and compared. As a result, Fig. 7 shows that CorrRISE consistently obtains more stable saliency maps than other approaches and can always precisely highlight the most similar regions between any given image pair. The adapted Grad-CAM and Grad-CAM++ methods produce less stable and meaningful explanation maps. LIME and RISE tend to allocate a very broad range of high-saliency pixels, making the importance map less precise. xFace and MinPlus achieve comparable results with CorrRISE. But the former fails in a difficult sample (row 4), and the latter does not exclude mask regions (row 5).

On the other hand, the samples in Fig. 3 and 7 are selected from diverse demographic groups, e.g. various genders, ethnicities, and ages. The visual results have validated that the CorrRISE method only depends on the decision of the verification system and shows no significant bias across demographic groups.

5.4. Quantitative Evaluation Results

This section reports the “Deletion” and “Insertion” metrics for a quantitative evaluation for general explainable

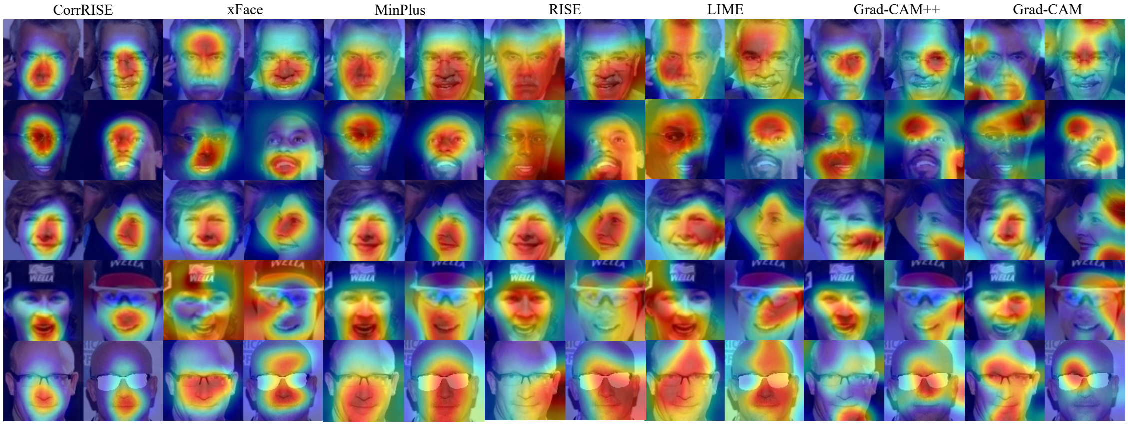

Figure 7. Visual results comparison among six saliency map-based explanation methods. The importance increases from blue to red color.

Table 1. Quantitative evaluation of saliency maps using proposed Deletion and Insertion metrics (%) on LFW, CPLFW, and Webface-OCC datasets, representing three different verification scenarios. **Del** (↓) refers to the Deletion metric, the smaller the better. **Ins** (↑) refers to the Insertion metric, the larger the better.

Methods	LFW		CPLFW		WebfaceOcc	
	Del	Ins	Del	Ins	Del	Ins
Grad-CAM [27]	50.65	65.00	40.15	56.77	35.33	69.67
Grad-CAM++ [5]	46.40	68.02	37.08	58.26	36.78	68.51
LIME [26]	38.48	78.72	28.83	70.59	24.04	82.15
RISE [24]	34.40	81.83	30.86	67.71	28.00	84.25
xFace [17]	35.21	79.95	32.34	65.69	27.72	81.17
MinPlus [21]	29.98	85.24	21.27	75.24	18.32	89.22
CorrRISE	**23.29**	**85.70**	**17.60**	**77.88**	**13.24**	**89.70**

face verification models. As described in Section 4, the metrics measure the changes in verification accuracy after modifying the input images according to the saliency map. Considering that some methods cannot deal with non-matching cases and for a fair comparison, evaluation is conducted based on the matching pairs of each dataset.

Table 1 shows the quantitative comparison among the state-of-the-art explanation methods with the deletion and insertion metrics in three typical face verification scenarios. In general, the reported metrics are consistent with the visual observations in Fig. 7 and those visual results in supplementary materials. For example, the adapted Grad-CAM and Grad-CAM++ show poorer performance than other SOTA explanation methods. It is notable that CorrRISE achieves much better quantitative results than a straightforward adaptation of RISE. The latter can only obtain comparable results with other adapted XAI methods such as LIME, demonstrating the advantage and value of our proposed CorrRISE method. Although the two XFR methods, xFace and MinPlus, present visually compelling results, quantitative metrics show that CorrRISE is capable of providing more precise saliency maps for all three datasets.

Table 2. Explainablity performance of CorrRISE tested on different state-of-the-art face recognition models. The verification accuracy (%) of FR models and two explainability metrics (%) for CorrRISE are reported.

FR Models	Acc (LFW)	Deletion (↓)	Insertion (↑)
ArcFace [7]	99.53	23.29	85.70
MagFace [20]	99.83	23.35	85.68
AdaFace [16]	99.82	23.31	85.63

CorrRISE is additionally applied to three different state-of-the-art FR models to show the validity and generalization ability of our proposed method. Table 2 shows that, when all the models achieve similar verification accuracy on LFW dataset, the generated saliency maps also report similar scores on "Deletion" and "Insertion" metrics, which proves that CorrRISE is model-agnostic.

6. Conclusion

In this work, the problem of explanation for face verification has been explored. Specifically, a new explanation framework for face verification was conceived. The proposed CorrRISE algorithm produces precise and insightful saliency maps to interpret the decision-making process of a deep FR model. Extensive experimental results in multiple scenarios show the advantage of our proposed method and its capability in analyzing the potential verification failures in challenging scenarios and, thereby provide insights into improving the current FR system. Meanwhile, a new evaluation methodology is designed, which offers a fair comparison among the saliency map-based explainable face verification methods and benefits future research in this area.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 5

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 1

[3] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016. 1, 2

[4] Gregory Castanon and Jeffrey Byrne. Visualizing and quantifying discriminative features for face recognition. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 16–23. IEEE, 2018. 2, 4

[5] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 1, 2, 5, 7, 8

[6] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017. 2

[7] 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. 5, 8

[8] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for content-based image retrieval. In *CVPR Workshops*, pages 95–98, 2019. 2

[9] Omar Elharrouss, Noor Almaadeed, and Somaya Al-Maadeed. Lfr face dataset: Left-front-right dataset for pose-invariant face recognition in the wild. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, pages 124–130. IEEE, 2020. 5

[10] 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. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[12] Bernease Herman. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*, 2017. 4

[13] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 440–450, 2022. 2, 4

[14] Baojin Huang, Zhongyuan Wang, Guangcheng Wang, Kui Jiang, Kangli Zeng, Zhen Han, Xin Tian, and Yuhong Yang. When face recognition meets occlusion: A new benchmark. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4240–4244. IEEE, 2021. 5

[15] 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. 5

[16] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18750–18759, 2022. 5, 8

[17] 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. 2, 5, 7, 8

[18] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9215–9223, 2018. 1

[19] Yu-Sheng Lin, Zhe-Yu Liu, Yu-An Chen, Yu-Siang Wang, Ya-Liang Chang, and Winston H Hsu. xcos: An explainable cosine metric for face verification task. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3s):1–16, 2021. 2

[20] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 5, 8

[21] Domingo Mery. True black-box explanation in facial analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1596–1605, 2022. 2, 3, 5, 7, 8

[22] Domingo Mery and Bernardita Morris. On black-box explanation for face verification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3418–3427, 2022. 2

[23] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017. https://distill.pub/2017/feature-visualization. 1

[24] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 2, 4, 5, 7, 8

[25] 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. 2

[26] 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. 2, 5, 7, 8

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2, 5, 7, 8

[28] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? *arXiv preprint arXiv:1611.07450*, 2016. 2

[29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2

[30] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 2029–2037. IEEE, 2019. 2

[31] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 248–263. Springer, 2020. 1, 2, 3

[32] Martin Winter, Werner Bailer, and Georg Thallinger. Demystifying face-recognition with locally interpretable boosted features (libf). In *2022 10th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2022. 2

[33] Zewei Xu, Yuhang Lu, and Touradj Ebrahimi. Discriminative deep feature visualization for explainable face recognition. *arXiv preprint arXiv:2306.00402*, 2023. 4

[34] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9348–9357, 2019. 2

[35] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 1

[36] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018. 1, 2, 4

[37] 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, 2018. 5

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1, 2