Explanation of Face Recognition via Saliency Maps

Yuhang Lu and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG), École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

ABSTRACT

Despite the significant progress in recent years, deep face recognition is often treated as a "black box" and has been criticized for lacking explainability. It becomes increasingly important to understand the characteristics and decisions of deep face recognition systems to make them more acceptable to the public. Explainable face recognition (XFR) refers to the problem of interpreting why a recognition model matches a probe face with one identity over others. Recent studies have explored use of visual saliency maps as an explanation mechanism, but they often lack a deeper analysis in the context of face recognition. This paper starts by proposing a rigorous definition of explainable face recognition (XFR) which focuses on the decision-making process of the deep recognition model. Based on that definition, a similarity-based RISE algorithm (S-RISE) is then introduced to produce high-quality visual saliency maps for a deep face recognition model. Furthermore, an evaluation approach is proposed to systematically validate the reliability and accuracy of general visual saliency-based XFR methods.

Keywords: Face Recognition, Explainability, Evaluation

1. INTRODUCTION

Thanks to the rapid development of deep learning, recent years have witnessed a breakthrough in various computer vision tasks, such as image classification, object detection, and face recognition. Deep face recognition has attracted worldwide attention in the past decade due to its remarkable performance and wide applications in multiple areas, such as access control, and video surveillance, to mention a couple. Despite their benefits, systems relying on face recognition also have the potential to endanger fundamental privacy and data protection rights, raising serious public concerns. Besides, the "black-box" nature of the deep learning-based systems is another barrier to the deployment of face recognition which is often criticized for its bias and lacking transparency and interpretability. Therefore, it is necessary to understand and explain the decisions made by deep face recognition technologies in order to further improve their performance and to make them more acceptable to the society at large.

A number of explanation techniques were first proposed as forms of explainable artificial intelligence (XAI) to better understand AI-based models. For computer vision tasks, various visual saliency map algorithms[1–7] have been introduced to highlight either the internal CNN layers or the important pixels of the input image that are relevant to the model's decision.

Explainable face recognition (XFR) is the problem of explaining how the face recognition model verifies a given pair of faces. Although numerous visual saliency algorithms have achieved impressive results in classification tasks, they cannot be directly applied to other image-understanding tasks due to a notable difference in internal model structure and the output format. To address this issue, some studies have attempted to adapt the existing explanation method to the face recognition task. Authors in[8,9] leveraged the contrastive excitation backpropagation (cEBP) technique[7] to localize important regions in the face. Mery et al.[10] adopted a similar idea of perturbing input images as in.[3,6] Winter et al.[11] applied an explainable boosting machine to face verification. Nevertheless, explaining a face recognition model not only refers to generating a saliency map but also involves an interpretation of why the model believes a certain pair of images is a better match than others and there is a lack of sufficient discussion regarding the latter in the literature. This paper first presents a

Further author information: (Send correspondence to the authors)
E-mail: yuhang.lu@epfl.ch, touradj.ebrahimi@epfl.ch

new definition of visual saliency-based explainable face recognition. Then a similarity-based RISE algorithm adapted from[6] is introduced. It provides insightful saliency explanations for the decision-making process of deep face recognition models. Moreover, to validate the reliability of the generated saliency maps, the paper further adapts and improves the classical "Deletion" and "Insertion" metrics to explainable face recognition tasks, providing a new benchmark for current XFR methods.

2. RELATED WORK

2.1 Face Recognition

In recent years, deep learning has revolutionized the field of face recognition and current deep face recognition techniques have shown impressive performance on different public benchmarks. The advances of such technologies mainly come from publicly available large-scale face datasets, powerful network architectures, and the evolution of training losses.

Early versions of deep face recognition systems were often built upon shallow network architectures, e.g. FaceNet[12] used GoogLeNet[13] and VGGFace[14] used VGGNet.[15] Then, ResNet[15] introduced residual learning to train very deep networks and has become a popular network architecture for current face recognition models.[16] further improved the ResNet by introducing a squeeze and excitation module, often denoted by SE-ResNet. While researchers first focused on investigating deeper backbone networks, the attention has gradually shifted to designing more discriminative loss functions. The most commonly used SoftMax loss[17] was extended with an angular margin to improve the intra-class compactness and inter-class discrepancy. SphereFace[18] first applied a multiplicative angular margin in the face recognition task and obtained better performance when compared to previous work. CosFace[19] introduced easier supervision by directly adding a cosine margin penalty to the target logit. Subsequently, ArcFace[20] proposed an additive angular margin loss to better separate the feature. More recently, the latest research, such as MagFace[21] and AdaFace,[22] created loss functions that are adaptive to image quality.

Despite the remarkable performance of current deep face recognition methods, they still suffer from a number of non-trivial scenarios according to recent studies,[23] such as partial occlusion, low-resolution, and head pose variation. It is necessary to have explainable tools that can provide insights into why a face recognition system fails in certain situations and how to take action to improve the current recognition process.

2.2 Visual Saliency Methods for Explainable Artificial Intelligence

Explainable artificial intelligence (AI) is one of the most important problems in machine learning, which aims at a better comprehension of the decisions made by AI-based models, such as those by deep neural networks. There are several taxonomies of explanations based on their scope and mechanism. For example, a main criterion is *intrinsic vs post-hoc*. Intrinsic methods refer to the machine learning models that are explainable by themselves, while post-hoc interpretability refers to explanation methods that are applied to a regular model after training.

Visual saliency algorithms have been widely employed to explain deep learning-based decision systems acting on images and are often used to provide post-hoc explanations for deep neural networks. A saliency map is itself an image, in which each pixel value represents its importance. In practice, a saliency map is modeled as a 2-D matrix which is of the same size as the image to be explained. The value of each pixel in the saliency map represents the importance of the corresponding pixel in the input image. The saliency map provides insight into the important regions of the input that are responsible for the model's final decision. In general, there are two types of methods for creating saliency maps.

The first group of methods backpropagates an importance score through the layers of the neural network from the model's output to the individual pixels in the input. This type of approach often requires access to the intrinsic architecture or gradient information of the deep model. The class activation maps (CAM) method[2] was one of the earliest works in this area, replacing the fully-connected layers with a global average pooling layer and obtaining the class-specific importance region by computing a weighted sum of the feature activation values. Grad-CAM[4] improved CAM by weighing the feature activation values with the class-specific gradient information that flows into the final convolutional layer of a CNN. Grad-CAM++[5] further extended the Grad-CAM by providing a better visual explanation of CNN model predictions. Layer-wise relevance propagation

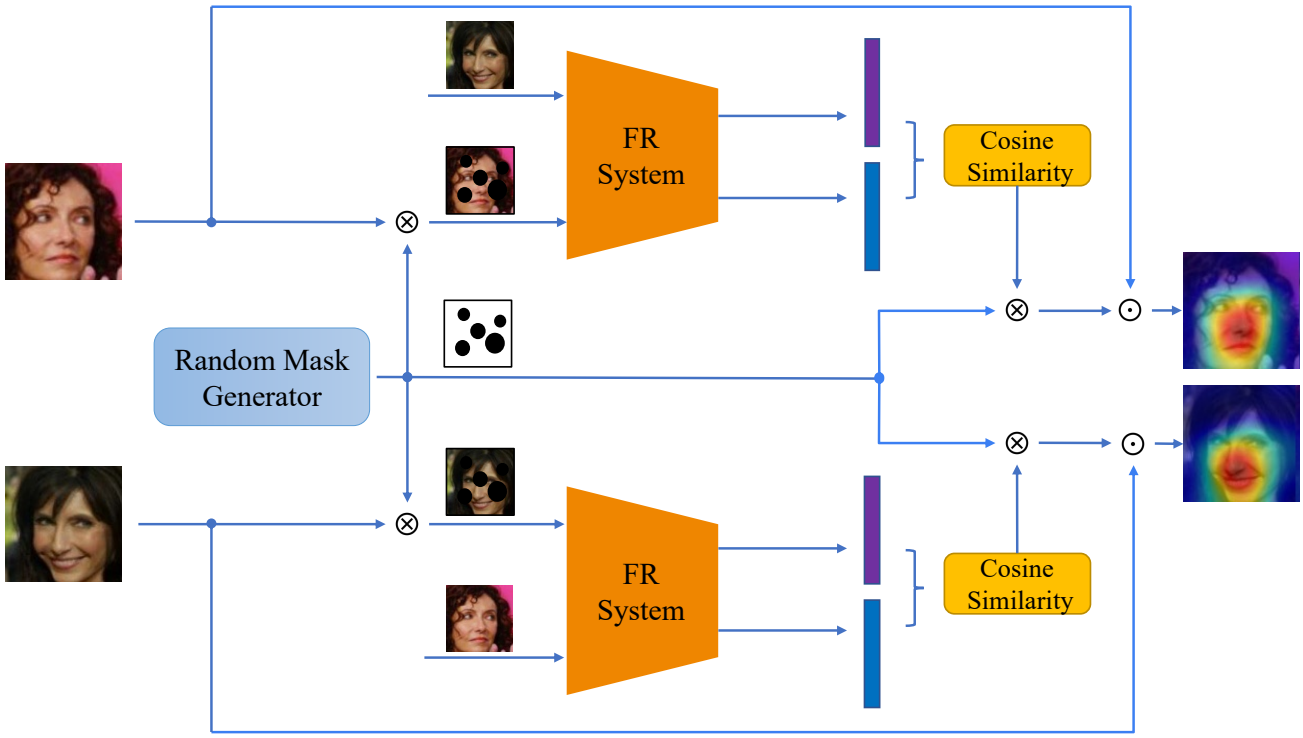

Figure 1: Workflow of the proposed S-RISE explanation method.

(LRP)[1] provides post-hoc explanation by decomposition and is capable of interpreting the decisions of complex neural networks. Specifically, it redistributes the prediction of the neural network backward until it assigns a relevance score to each input image pixel.

An alternative method performs random perturbations on the image, e.g. noise, occlusion, etc, and determines the importance region by observing the impact of such perturbation on the model's output prediction. For instance, Ribeiro et al.[3] 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 manner. Randomized input sampling for explanation of black-box models (RISE)[6] algorithm applies random binary masks 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. Taking advantage of the simple mechanism, it has been adapted to explain other vision tasks.[24] While most of the popular visual saliency explanation methods are initially developed for image classification tasks, there has been an increasing demand for creating explanations for other image-understanding tasks such as object detection.[24,25]

2.3 Visual Saliency Methods for Explainable Face Recognition

In face recognition, earlier work[8] adapted several saliency map creation algorithms from classification tasks, such as Grad-CAM,[4] and Guided Grad-CAM,[26] for the face recognition task and compared their performance using a metric called the "hiding game". Williford et al.[9] first proposed a comprehensive benchmark for explainable face recognition and additionally adopted subtree excitation backprop (EBP) method[7] to produce explainable saliency maps. Mery et al.[10] introduced six different perturbation-based methods to create saliency maps to explain the face verification model without manipulating the model. Some studies have adopted intrinsic approaches and attempted to increase explainability by introducing external modules to a face recognition system. Yin et al.[27] designed a feature activation diverse loss to encourage learning more interpretable face representations. Lin et al.[28] proposed a learnable module that can be integrated into face recognition models and generates meaningful similarity maps.

3. PROPOSED METHODS

3.1 Definition for Explainable Face Recognition

In the context of face recognition, the deep model predicts whether a pair of face images belong to the same identity. Ideally, an explainable face recognition system gives a visual interpretation of why the model believes the given pair of faces is a match or non-match. Existing study[9] has explored a similar idea of explaining how a deep face recognition model matches faces. It leverages a triplet of faces, i.e. probe, mate, and nonmate, to provide a deeper explanation for the relative importance of facial regions. Notably, probe refers to the query image to be verified, mate is the image from the same subject as the probe, and nonmate is the image from another subject. More specifically, explainable face recognition was defined as a way to highlight the regions of the probe image that can maximize the similarity to the mate image and meanwhile minimize the similarity to the nonmate.

However, there is a drawback in this definition of explanation, as the most similar regions between the probe and the mate images are not necessarily the least similar regions between the probe and the nonmate. In fact, a face recognition system makes decisions by comparing a predefined threshold with the similarity score between two images instead of three, which means in the triplet, the decision-making process for each pair of images is independent.

This paper proposes a more rigorous definition of explainable face recognition, which preserves the idea of triplet images but disentangles the matching and non-matching pairs. Given a randomly selected triplet of probe, mate, and nonmate images, denoted by $\{I_p, I_m, I_n\}$, feeding into a face recognition system respectively, the explanation method should produce the corresponding saliency maps for the $\{I_p, I_m\}$ and $\{I_p, I_n\}$ pairs, which answer to the following questions.

- Which regions in the $\{I_p, I_m\}$ image pairs are the most similar to the FR system?
- Which regions in the $\{I_p, I_n\}$ image pairs are the most similar to the FR system?
- Why the FR system believes that $\{I_p, I_m\}$ pair is a better matching than $\{I_p, I_n\}$?

3.2 Similarity-based Randomized Input Sampling for Explanation (S-RISE)

The previous section gives a new definition to the problem of explainable face recognition by stipulating that an explanation method should both produce proper saliency maps for critical regions of the input image and interpret the decision of the recognition model. This paper proposes a new model-agnostic explanation method for face recognition following this new definition of XFR.

The RISE method explains a classification model by leveraging the categorical output probability of the classifier as the weight to aggregate the final saliency map. It is capable of providing more precise explanation maps when compared to other model-agnostic methods. Despite being useful in principle, existing explainability tools for other image understanding tasks cannot be directly applied to the face recognition task. Because the decision-making process of a face recognition system mainly involves deep face representation extraction and similarity calculation between at least two images. This paper proposes a Similarity-based RISE algorithm (S-RISE) that leverages the similarity score as weights for the masks and provides explanation saliency maps without accessing the internal architecture or gradients of the face recognition system.

Figure 1 depicts an overview of the proposed S-RISE algorithm and it takes a pair of probe and mate input images as examples. In general, given a pair of images $\{I_p, I_m\}$, a mask generator will first randomly produce N masks, denoted by $M = \{M_i, i = 1, ..., N\}$. Each mask M_i will be applied to the input image, e.g. I_p. The masked $I_p \odot M_i$ and unmasked I_m are then fed into the face recognition model respectively to capture the deep face representation. Afterward, the cosine similarity is computed and used to weigh the corresponding mask. After iterating all the masks, the final saliency map H_p for I_p is the weighted combination of the generated masks. More specifically, the proposed S-RISE algorithm comprises the following two pivot steps, i.e. mask generation and saliency map generation.

3.2.1 Mask Generation

The conventional RISE algorithm for image classification tasks aims at generating multiple random non-binary masks. Basically, the authors propose to first sample small binary masks and then upsample them to a larger resolution with bilinear interpolation, after which the masks have values between [0,1]. Here we propose to simplify the process by directly generating multiple small Gaussian-distributed patches in random locations. In practice, the mask generation process is as follows.

1. Initialize the parameters of the mask generator, i.e. the total number of masks N, and the number and size of the Gaussian kernels in each mask.
2. For each mask M_i, sample multiple patches that follow Gaussian distribution at random locations.
3. Repeat step 2 to get N different masks $\{M_i, i = 1, ..., N\}$.

3.2.2 Similarity-based Saliency Map Generation

This subsection describes in more details how the S-RISE algorithm produces saliency maps for a triplet input. The S-RISE method is designed to explain the predictions between every independent pair of images due to the special decision-making process of the face recognition system. Thus, the triplet will be first divided into two groups, i.e. matching pair $\{I_p, I_m\}$ and non-matching pair $\{I_p, I_n\}$, and the S-RISE algorithm will explain their corresponding predictions, respectively. In general, the mask generator randomly samples a fixed number of masks first for the matching pair and the S-RISE algorithm starts the iteration to compute the saliency map. The same steps are repeated for the non-matching pair. In the end, all the produced visual explanation heatmaps will be normalized. The following operation describes in more details the procedures used in the proposed S-RISE algorithm.

1. Split the triplet input into $\{I_p, I_m\}$ and $\{I_p, I_n\}$ pairs.
2. Sample N masks, $M = \{M_i, 1 \leq i \leq N\}$, using the mask generator. Apply M_i to the I_p and I_m separately.
3. Iterate i from 1 to N:
   - 3.1. Forward the masked probe image $I_p \odot M_i$ and unmasked mate image I_m into the network for feature extraction, calculate similarity score s_i^p between the deep features.
   - 3.2. Forward the masked mate image $I_m \odot M_i$ and unmasked probe image I_p into the network for feature extraction, calculate similarity score s_i^m between the deep features.
4. Compute the weighted sum of masks M_i with respect to the calculated similarity score s_i^p to obtain a saliency map for the probe image $H_p = \sum_{i=1}^{N} s_i^p M_i$.
5. Compute the weighted sum of masks M_i with respect to the calculated similarity score s_i^m to obtain a saliency map for the mate image $H_m = \sum_{i=1}^{N} s_i^m M_i$
6. Repeat steps 2-5 for input pair $\{I_p, I_n\}$
7. Normalize the saliency maps for both matching and non-matching pairs and obtain $\{H_p, H_m\}$ and $\{H_p, H_n\}$.

In practice, it is notable that the masks are only applied to one image at the same time during Step 3. The reason is that, due to the special decision-making process in the face recognition task, two irrelevant but masked images can be easily treated as matching pairs, which interferes calculation of similarity scores.

3.3 Evaluation Methodology

The importance of rigorous evaluation methodologies has been overlooked in the field of explainable artificial intelligence and only a few metrics have been designed for visual saliency explanation methods. In the context of explainable image classification tasks, Petsiuk et al.[6] insert or delete salient pixels from the input image and measure the change in the output classification probability.[29] adopted the same "Deletion" and "Insertion" metrics to image retrieval task. In face recognition,[8] quantifies 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 existing work[30] shows that it is not able to differentiate the well-performed explanation methods.

In order to systematically validate the reliability of the saliency maps generated by XFR methods, this paper adopted the conventional "Deletion" and "Insertion" evaluation metrics and further improved them to better fit the explainable face recognition framework. The main insight is that the explanation saliency map is expected to precisely highlight the most important regions of the face with the smallest number of pixels for the face recognition model to take a correct decision.

In general, the Deletion and Insertion metrics measure how fast the similarity between two faces drops/rises to a threshold value after removing/adding saliency pixels from them. More specifically, the deletion process starts with original images, and the pixels with the highest saliency values are sequentially removed and replaced with a constant value. After removing each pixel, the similarity score is re-calculated until it is lower than a predefined threshold. On the contrary, the insertion process starts with the constant value, and the most critical pixels in the image sorted by the saliency map are added to the plain image. The similarity score is re-calculated each time after adding one pixel until it is larger than the threshold. The number of pixels deleted from or added to the image is accumulated until the recognition model changes the decision. Overall, the deletion and insertion metrics are defined as $\frac{\#Removed\ pixels}{\#All\ pixels}$ and $\frac{\#Added\ pixels}{\#All\ pixels}$. In practice, directly removing pixels from an image alters the original distribution and can eventually affect recognition results.[8,31] Hence, the constant value above is set as the mean value of the specific image.

4. EXPERIMENTAL RESULTS

4.1 Implementation Details

4.1.1 Face Recognition Model Setup

This paper utilizes the ArcFace[20] face recognition approach, with a backbone network of ResNet-50,[32] for the explanation experiments, because this is the most commonly employed combination in both industry and academia. The architecture of the face recognition model remains unchanged during the explanation. The face recognition model is trained with the cleaned version of MS1M dataset,[20] which is composed of approximately 5.1M face images belonging to 93K identities.

4.1.2 Explanation Model Setup

The proposed S-RISE explainability method does not require any training or access to the internal architecture of the face recognition model. During the explanation process, the number of iterations in the explanation process is set to 500 by default. For each mask, there are 100 patches and each patch is generated by a Gaussian kernel which is configured with kernel size $ks = 40$ and $\sigma = 7$.

4.1.3 Dataset

To validate the effectiveness of the proposed explainability approach, experiments have been carried out on testing images from several popular databases. First, the triplets selected for visual results demonstration are randomly sampled from LFW,[33] CPLFW,[34] and Webface-Occ[35] datasets, which cover the standard, pose-variant, and occluded face recognition scenarios. Then, a small subset of the LFW dataset is sampled, which comprises 50 triplets of faces. This subset is used for the quantitative evaluation purpose.

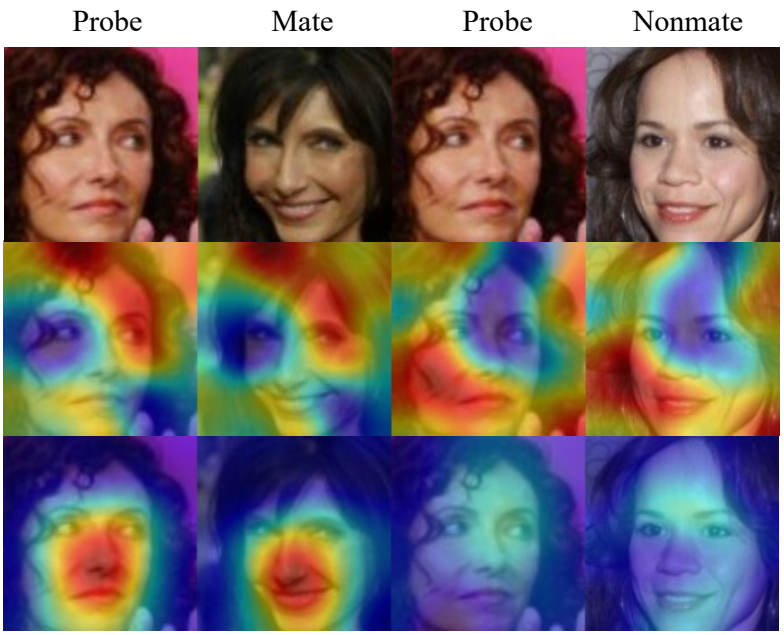

Figure 2: Sanity check for the S-RISE explanation method. The second row is the generated explanation heatmap for a CNN model with randomized parameters, while the third row is for a normal face recognition system.

4.1.4 Preprocessing

All the test images strictly follow the same preprocessing steps as the training data. The MTCNN[36] is first applied to detect faces and landmark points. All images are then cropped, aligned, and resized to the size of 112x112 pixels.

4.2 Sanity Check for the Proposed Explanation Method

A recent study[37] has questioned the validity of saliency methods that some of the produced explanation heatmaps can be independent both of the model and of the data generating process. They propose a model parameter randomization test for a sanity check, where the weights of a deep neural network are randomly initialized before applying the explanation method. In the context of this paper, a saliency method can also provide some visually compelling results by simply applying a heatmap that concentrates on the center of the face and claim it to be an attention map without analyzing the behavior of the recognition model. Therefore, a similar idea for the sanity check is adopted. Specifically, the S-RISE algorithm is used to explain the decision of a face recognition model, whose parameters are randomly initialized.

The explanation results in the second row of Figure 2 show that the random parameters of the deep model will result in nonsense saliency maps, which validates that the proposed S-RISE algorithm fully relies on the trained recognition model and is capable of producing meaningful interpretations.

4.3 Visual Results of Saliency Maps

This section presents the visual results of the saliency map generated by the proposed S-RISE algorithm. The experiments have been conducted under three scenarios, namely standard verification, incorrect verification, and self-occlusion cases. In the standard verification scenario, Figure 3 presents the visual explanations for the predictions that the face recognition model correctly makes. As a result, the produced saliency map properly highlights the regions between the matching pairs that the FR model believes are very similar. As for the probe and nonmate pairs, the heatmaps also represent similar regions but they are much shallower, indicating lower similarities between them when compared to the matching pairs.

While the current deep face recognition model generally achieves high prediction accuracy with high similarity scores in most standard scenarios, it mistakenly identifies two subjects as the same person in some cases. In the

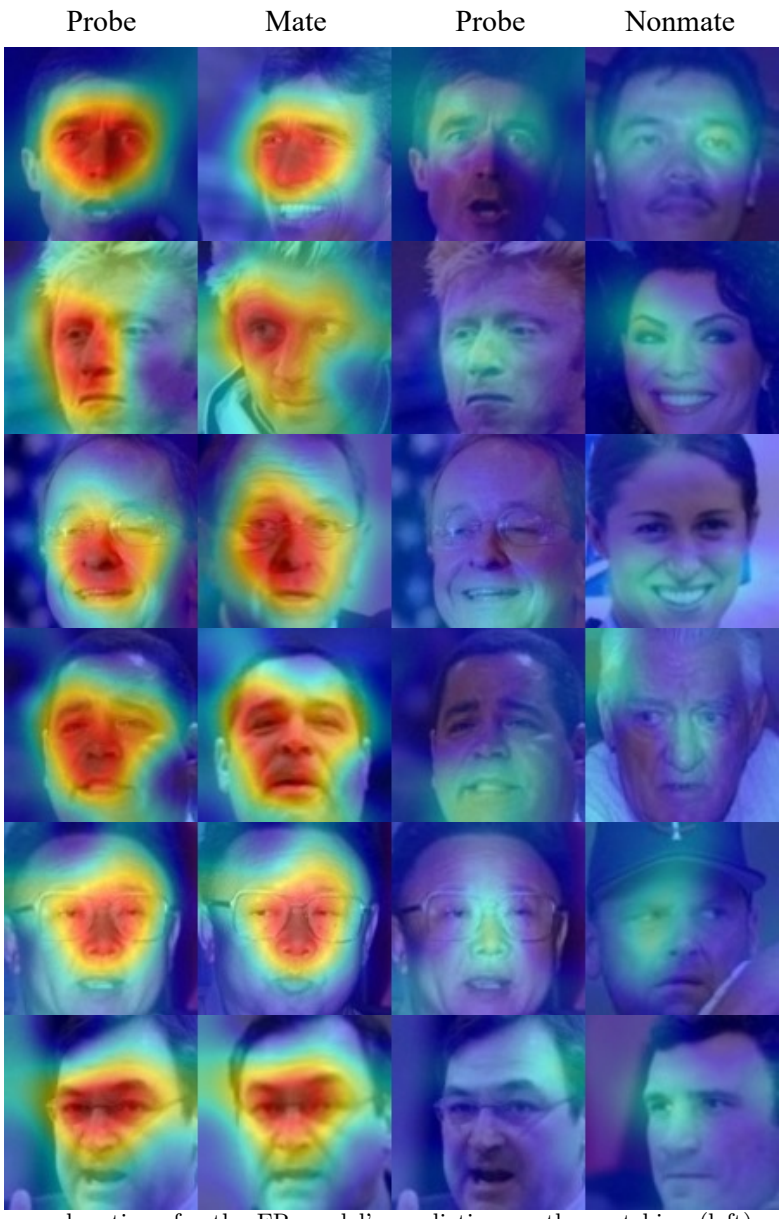

Figure 3: Saliency map explanations for the FR model's prediction on the matching (left) and non-matching (right) image pairs.

second testing scenario, the S-RISE algorithm is applied to a manually selected triplet that the face recognition system fails to differentiate, see Figure 4. According to the explanation heatmap, although it recognizes the probe-mate pair with high similarity scores, it also allocates very high saliency values to the eye and mouth regions to the probe-nonmate pairs, which explains why the FR model mistakenly verifies the non-matching pair as from the same person.

To further validate the effectiveness of the proposed S-RISE explanation method, an additional test has been performed with self-occluded faces. Studies[23] have shown that the current deep face recognition model is capable of identifying partially occluded faces despite allocating relatively low similarity scores. In this context, an ideal explanation method should only spotlight the non-occluded regions of the face and neglect the masked regions. In this experiment, the non-matching face in a triplet is replaced by an occluded image from the same subject as the probe face. As presented in Figure 5, the S-RISE algorithm explains that the FR model manages to verify

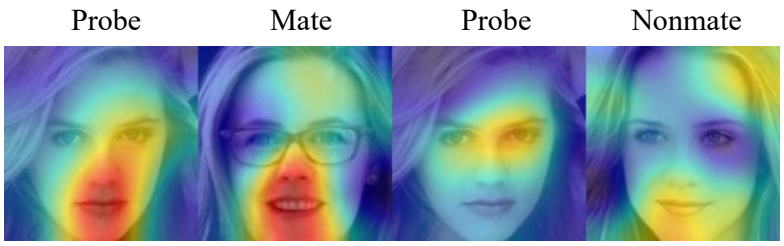

Figure 4: Saliency map explanation on failed predictions of the face recognition model.

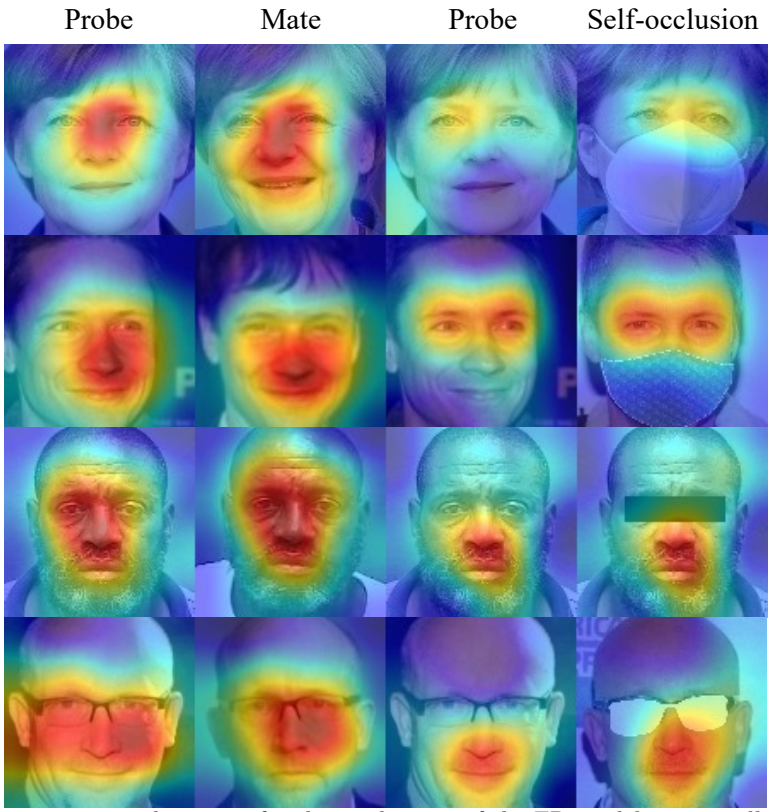

Figure 5: Saliency map explanations for the predictions of the FR model on partially-occluded faces.

them through the eye regions when images are occluded by facial masks, and through mouth and nose areas when masked by sunglasses.

4.4 Quantitative Evaluation

This section reports the deletion and insertion metrics as quantitative evaluations for the proposed S-RISE explanation model. The experiments are conducted on a subset of the LFW dataset. In general, the metrics measure the percentage of modified pixels in order to change the decision of the FR model. The smaller, the more accurate the explanation saliency map.

To validate the effectiveness of the proposed evaluation method, it has been applied to the proposed S-RISE algorithm which is under different configurations in terms of iterations. Figure 6 shows that a small number of iterations, i.e. 10, will result in poor saliency maps. Simultaneously, the quantitative evaluation method gives poor scores to this misconfigured explanation model, as shown in Table 1. On the other hand, the two examples on the right side of Figure 6 show that the S-RISE converges after around 500 iterations, this can also be validated by the metrics reported in the table. While RISE-based algorithms are known for introducing

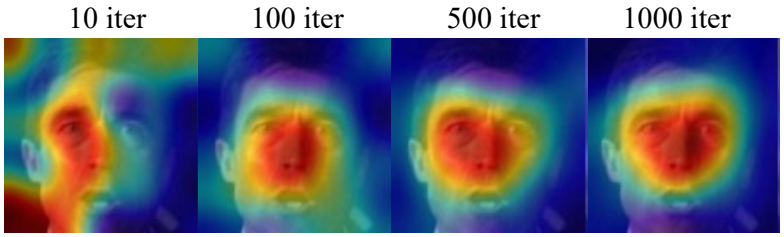

Figure 6: Saliency map generated by S-RISE algorithm with different iteration configurations.

Table 1: Quantitative evaluation of saliency maps using proposed Deletion and Insertion metrics.

Methods	Iterations	Deletion	Insertion	Average
S-RISE	10	0.4466	0.3582	0.4024
	100	0.2617	0.1983	0.2300
	500	0.2071	0.1459	0.1765
	1000	0.2077	0.1384	0.1731

inevitable randomness to final predictions, it is notable that the proposed S-RISE is able to provide stable and accurate saliency maps for face recognition models given sufficient iterations.

5. CONCLUSION

In this paper, a new explainable face recognition framework was conceived, implemented, tested and validated. The proposed S-RISE algorithm is capable of producing insightful saliency maps to interpret the decision of a deep face recognition system. Extensive visual results of saliency maps have demonstrated the effectiveness of the method. Furthermore, two evaluation metrics are introduced in this work to measure the quality of the saliency maps generated by the S-RISE algorithm. In the future, the evaluation method can serve as a public benchmark for general visual saliency map-based XFR methods.

ACKNOWLEDGMENTS

The authors acknowledge support from CHIST-ERA project XAIface (CHIST-ERA-19-XAI-011) with funding from the Swiss National Science Foundation (SNSF) under grant number 20CH21 195532.

REFERENCES

[1] Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W., "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*], 63–71, Springer (2016).

[2] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Learning deep features for discriminative localization," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 2921–2929 (2016).

[3] Ribeiro, M. T., Singh, S., and Guestrin, C., "" 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*], 1135–1144 (2016).

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[5] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N., "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in [*2018 IEEE winter conference on applications of computer vision (WACV)*], 839–847, IEEE (2018).

[6] Petsiuk, V., Das, A., and Saenko, K., "Rise: Randomized input sampling for explanation of black-box models," *arXiv preprint arXiv:1806.07421* (2018).

[7] Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S., "Top-down neural attention by excitation backprop," *International Journal of Computer Vision* **126**(10), 1084–1102 (2018).

[8] Castanon, G. and Byrne, J., "Visualizing and quantifying discriminative features for face recognition," in [*2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*], 16–23, IEEE (2018).

[9] Williford, J. R., May, B. B., and Byrne, J., "Explainable face recognition," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*], 248–263, Springer (2020).

[10] Mery, D. and Morris, B., "On black-box explanation for face verification," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 3418–3427 (2022).

[11] Winter, M., Bailer, W., and Thallinger, G., "Demystifying face-recognition with locally interpretable boosted features (libf)," in [*2022 10th European Workshop on Visual Information Processing (EUVIP)*], 1–6, IEEE (2022).

[12] Schroff, F., Kalenichenko, D., and Philbin, J., "Facenet: A unified embedding for face recognition and clustering," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 815–823 (2015).

[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A., "Going deeper with convolutions," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 1–9 (2015).

[14] Parkhi, O. M., Vedaldi, A., and Zisserman, A., "Deep face recognition," (2015).

[15] Simonyan, K. and Zisserman, A., "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

[16] Hu, J., Shen, L., and Sun, G., "Squeeze-and-excitation networks," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7132–7141 (2018).

[17] Liu, W., Wen, Y., Yu, Z., and Yang, M., "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295* (2016).

[18] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L., "Sphereface: Deep hypersphere embedding for face recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 212–220 (2017).

[19] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W., "Cosface: Large margin cosine loss for deep face recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 5265–5274 (2018).

[20] Deng, J., Guo, J., Xue, N., and Zafeiriou, S., "Arcface: Additive angular margin loss for deep face recognition," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 4690–4699 (2019).

[21] Meng, Q., Zhao, S., Huang, Z., and Zhou, F., "Magface: A universal representation for face recognition and quality assessment," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 14225–14234 (2021).

[22] Kim, M., Jain, A. K., and Liu, X., "Adaface: Quality adaptive margin for face recognition," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 18750–18759 (2022).

[23] Lu, Y., Barras, L., and Ebrahimi, T., "A novel framework for assessment of deep face recognition systems in realistic conditions," in [*2022 10th European Workshop on Visual Information Processing (EUVIP)*], 1–6, IEEE (2022).

[24] Petsiuk, V., Jain, R., Manjunatha, V., Morariu, V. I., Mehra, A., Ordonez, V., and Saenko, K., "Black-box explanation of object detectors via saliency maps," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 11443–11452 (2021).

[25] Fong, R. C. and Vedaldi, A., "Interpretable explanations of black boxes by meaningful perturbation," in [*Proceedings of the IEEE international conference on computer vision*], 3429–3437 (2017).

[26] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in [*Proceedings of the IEEE international conference on computer vision*], 618–626 (2017).

[27] Yin, B., Tran, L., Li, H., Shen, X., and Liu, X., "Towards interpretable face recognition," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 9348–9357 (2019).

[28] Lin, Y.-S., Liu, Z.-Y., Chen, Y.-A., Wang, Y.-S., Chang, Y.-L., and Hsu, W. H., "xcos: An explainable cosine metric for face verification task," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **17**(3s), 1–16 (2021).

[29] Hu, B., Vasu, B., and Hoogs, A., "X-mir: Explainable medical image retrieval," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 440–450 (2022).

[30] Xu, Z., Lu, Y., and Ebrahimi, T., "Discriminative deep feature visualization for explainable face recognition," *arXiv preprint arXiv:2306.00402* (2023).

[31] Gomez, T., Fréour, T., and Mouchère, H., "Metrics for saliency map evaluation of deep learning explanation methods," in [*International Conference on Pattern Recognition and Artificial Intelligence*], 84–95, Springer (2022).

[32] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 770–778 (2016).

[33] Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E., "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).

[34] Zheng, T. and Deng, W., "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).

[35] Huang, B., Wang, Z., Wang, G., Jiang, K., Zeng, K., Han, Z., Tian, X., and Yang, Y., "When face recognition meets occlusion: A new benchmark," in [*ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 4240–4244, IEEE (2021).

[36] Zhang, K., Zhang, Z., Li, Z., and Qiao, Y., "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters* **23**(10), 1499–1503 (2016).

[37] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B., "Sanity checks for saliency maps," *Advances in neural information processing systems* **31** (2018).