2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Explainable Face Verification via Feature-Guided Gradient Backpropagation

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

***Abstract*— Recent years have witnessed significant advancement in face recognition (FR) techniques, with their applications widely spread in people's lives and security-sensitive areas. There is a growing need for reliable interpretations of decisions of such systems. Existing studies relying on various mechanisms have investigated the usage of saliency maps as an explanation approach, but suffer from different limitations. This paper first explores the spatial relationship between face image and its deep representation via gradient backpropagation. Then a new explanation approach FGGB has been conceived, which provides precise and insightful similarity and dissimilarity saliency maps to explain the "Accept" and "Reject" decision of an FR system. Extensive visual presentation and quantitative measurement have shown that FGGB achieves comparable results in similarity maps and superior performance in dissimilarity maps when compared to current state-of-the-art explainable face verification approaches.**

I. Introduction

Over the past decades, the accuracy of Face Recognition (FR) systems has been boosted due to the advanced technologies based on deep convolutional neural networks (DCNNs) [9], [11], [25], [27] and large-scale face datasets [3], [8], [34]. FR technology has become an increasingly important application, widely used in our daily lives and even security-critical applications, such as identity checks and access control. However, the DCNN-based FR systems often involve complicated and unintuitive decision-making processes, making it difficult to interpret or further improve them. To address this problem, significant efforts have been made with the objective of enhancing the transparency and interpretability of learning-based face recognition systems.

More recently, saliency algorithms have become a more intuitive way of explanation for general vision models by producing heat maps highlighting regions of the input image responsible for the model's output decision. These techniques were primarily developed for explainable image classification tasks [1], [4], [6], [21], [23], [24], [33], with a few addressing other explanation problems in image retrieval [10], object detection [22], etc. However, similar explanation algorithms for face recognition models are still under-explored, mainly due to the unique output formats and decision-making process. This paper aims to develop advanced saliency explanation algorithms for one of the most crucial problems in face recognition, i.e., eXplainable Face Verification (XFV), which essentially studies how a deep FR model matches a given facial image over another.

Unlike common image classification models that often produce categorical outputs, a deep face verification system takes a pair of face images as input. It first extracts deep representation for inputs and then calculates the cosine similarity between two face embeddings. The decision is made by comparing the similarity score with a predefined threshold. Some saliency-based explanation algorithms relying on different principles have been proposed to increase the explainability of the face verification process. For example, Lin et al. [18] plugged an external attention module to produce explainable heat maps. [16], [19], [20] applied random perturbations to the input face images and generated saliency maps by analyzing their impact on the verification output. In [28], the contrastive excitation backpropagation method is used to identify the relevant salient regions on the face images.

However, current explainable face verification approaches show some limitations. First, some methods [18], [20] only provide explanations when the FR model makes an "Accept" decision by presenting a similarity map while neglecting the reason for "Reject" decisions. Secondly, many popular XFV methods [16], [19], [20] are not efficient enough and take much longer time than the verification process, hindering further practical deployment. Saliency algorithms relying on gradient backpropagation are more efficient solutions, but they often suffer from the gradient fluctuation and vanishing problem, leading to noisy and inaccurate saliency maps. To address all these issues, this paper proposes a new Feature-Guided Gradient Backpropagation (FGGB) method that provides precise and efficient explanation saliency maps for arbitrary FR systems. The proposed FGGB method produces both similarity and dissimilarity maps between given input images. In practice, the former leads to the explanation for "Accept" decisions and the latter for "Reject" decisions. Moreover, FGGB performs gradient backpropagation at feature level instead of from final scores to explore the spatial relationship between the input image and its corresponding deep feature, followed by a new saliency map generation approach that prevents the noisy gradients problem.

II. Related Work

In general, most saliency-based explanation methods can be categorized into three groups based on their mechanisms. The first group aims to modify the internal architecture of the deep neural network to gain explainability. For example,

*Equal contribution

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

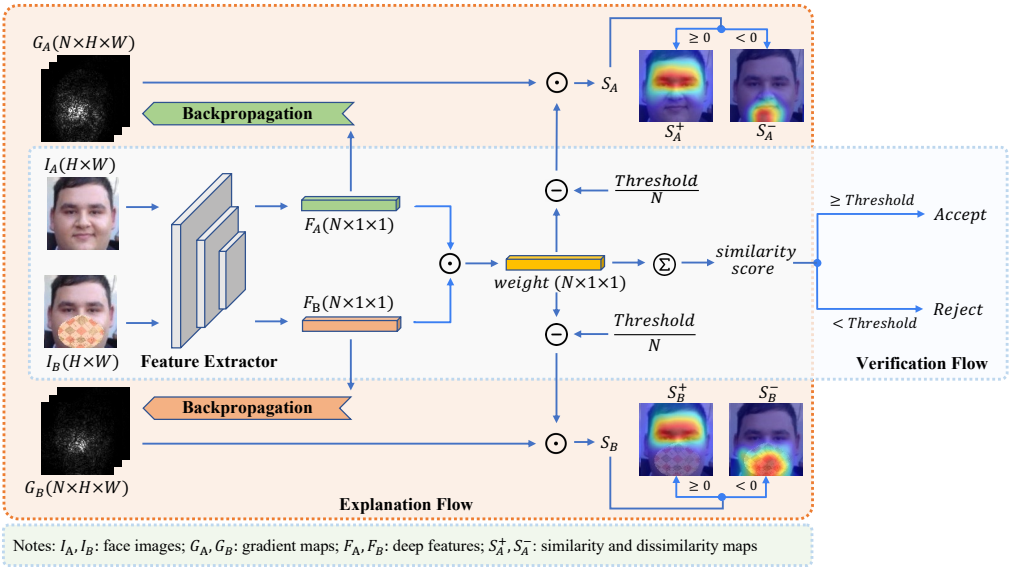

Fig. 1: Workflow of the proposed Feature-Guided Gradient Backpropagation method. The similarity and dissimilarity maps are calculated respectively given an arbitrary input face pair.

CAM [33] modified the last layer of the network. GAIN [17] integrated learnable modules into the training process to produce attention maps. In XFV, xCos [18] added a learnable attention module to the end of the verification pipeline, and Xu et al. [29] trained a deep FR model together with a face reconstruction network to preserve spatial information in the face representation. However, these methods all require retraining the entire face recognition model to learn more explainable deep representation and are often not applicable to systems that have already been deployed in the real world.

The second category is so-called perturbation-based methods, which determine the salient regions by observing the effect of a perturbation on the model's output. These methods work independently from the internal status of the deep neural network and offer "black-box" interpretation. The idea has been popular among recent XFV methods [2], [16], [19], [20]. For example, Mery [20] proposed to remove or aggregate different parts of images and highlight the most relevant parts for the verification process. Lu et al. [19] applied random masks to the input images and calculated both similarity and dissimilarity saliency maps through a correlation module. While these methods offer accurate and interpretable saliency maps, the perturbation-based mechanism generally lacks efficiency because they are obligated to run many iterations to guarantee a stable outcome.

Gradient backpropagation-based methods are typically more efficient solutions. In explainable image classification tasks, [30] calculated the derivatives of the categorical output with respect to the input image to identify the salient pixels on the image. In XFV, Huber et al. [14] backpropagated the cosine similarity score between two face images to obtain saliency maps that indicate similar and dissimilar regions. However, the outcomes of this type of method tend to be noisy. SmoothGrad [26] for classification task sharpened saliency maps by initiatively adding noise and averaging all the resulting gradient maps. Our method takes an alternative approach to resolve this problem, by backpropagating the gradients at feature level and re-weighting the gradient maps according to the importance of each feature channel.

III. Proposed Method

A. Problem Statement

In principle, a face verification system makes two types of decisions, i.e., "Accept" and "Reject". This paper offers interpretations from the user's perspective and explains why the face verification system believes the given pair of facial images are matching (Accept) or non-matching (Reject). More specifically, our saliency algorithm aims to provide similarity maps for acceptance decisions and dissimilarity maps for rejection decisions.

B. Feature-Guided Gradient Backpropagation (FGGB)

This paper proposes to leverage the gradient backpropagation method to calculate the saliency maps. Previous work [14] added a cosine similarity layer and directly backpropagated the output similarity score between two input face images throughout the verification model to get a saliency map, indicating which pixels contribute to the decision. However, a limitation of conventional gradient-based algorithms is that they identify salient pixels simply by observing raw gradient values, but the derivative of the output score may fluctuate sharply at small scales and even disappear during the backpropagation [26], leading to visually noisy saliency maps. In general, a deep face verification system relies on the direct comparison of the distance between two deep face features, and the most discriminative feature channels often dominate the final decision. The gradient value of the most important feature channel can possibly vanish due to the fluctuation, resulting in less accurate saliency explanations.

This work conceives a simple but effective propagation scheme and saliency map generation algorithm to resolve this issue, see Fig. 1. In specific, we perform gradient backpropagation from the deep feature level in a channel-wise manner rather than from the final score, and obtain multiple gradient maps. Instead of converting them to a saliency map, they are used to explore the spatial relationship between the image and its deep feature, because each gradient map propagated from an individual feature channel will spotlight certain pixels on the face image. Finally, the gradient maps are normalized and weight-summed by the channel-wise cosine similarity between two deep face representations, which guarantees high and stable saliency value for the most discriminative feature channels.

Given probe and gallery images $\{I_A, I_B\}$, the FR model extracts their deep face representations, each of dimension N, denoted by $\{F_A, F_B\}$. In the following, we explain the construction of similarity and dissimilarity saliency maps S_A^+ and S_A^- using the FGGB method in detail. That of S_B^+ and S_B^- follows similarly. The proposed method consists of two phases, (i) gradient backpropagation and (ii) saliency map generation.

In phase (i), one first backpropagates gradient from each channel of feature F_A and construct N gradient maps $G_A = \{G_A^k : k = 1, \ldots, N\}$ as

$$G_A^k = \frac{\partial F_A^k}{\partial I_A}, \tag{1}$$

where ∂F_A^k represents the derivative of the k-th dimension of feature F_A. Then, one performs normalization to each G_A^k to mitigate the impact of local variations (e.g., vanishing gradient during partial derivatives) and produces

$$\tilde{G}_A^k = \frac{|G_A^k|}{\|G_A^k\|}, \tag{2}$$

where $|\cdot|$ denotes the absolute value and $\|\cdot\|$ denotes the Frobenius norm of a matrix.

In phase (ii), all normalized gradient maps are accumulated to generate the saliency maps. First, a weight vector is defined to be the channel-wise cosine similarity between $\{F_A, F_B\}$:

$$weight = \frac{F_A \odot F_B}{\|F_A\|\|F_B\|}, \tag{3}$$

where $\odot$ denotes the elementwise product and $\|\cdot\|$ denotes the L_2 norm.

Then, features with large cosine similarity values will contribute to the similarity map and otherwise dissimilarity map. More specifically, one subtracts the decision threshold from the cosine similarity vector $weight$ and computes the saliency map S_A as the weighted sum of the gradient maps,

$$S_A = \sum_{k=1}^{N} \tilde{G}_A^k \cdot \Big(weight_k - \frac{threshold}{N}\Big), \tag{4}$$

which is then decomposed into similarity and dissimilarity maps

$$S_A^+ = S_A[S_A \geqslant 0], \quad S_A^- = S_A[S_A < 0]. \tag{5}$$

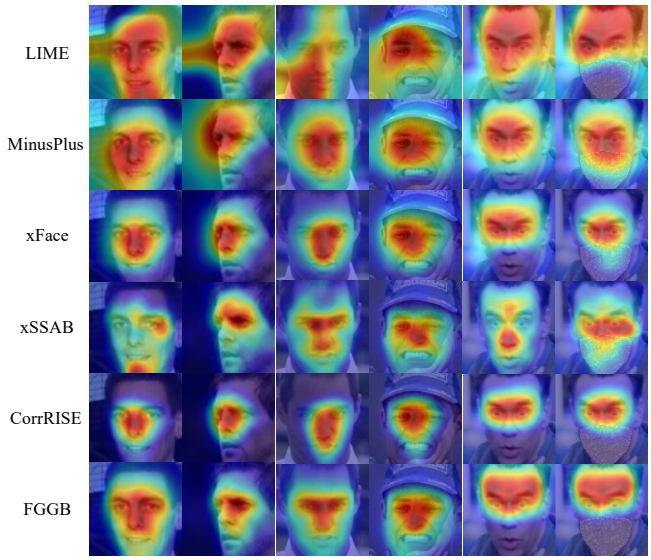

Fig. 2: Visual comparison of similarity maps generated by FGGB and five other XFV methods based on decisions of ArcFace model. Every two columns represent a pair of genuine faces. The saliency value increases from blue to red color.

IV. Experimental Results

A. Implementation Details

The proposed FGGB method is based on backpropagation and does not rely on any parameter or specific network architecture. As for deep face recognition models, this paper first conducts experiments on the popular ArcFace [7] model with iResNet-50 [11] backbone. To show the generalization ability of FGGB across various face recognition models, its explainability performance has been tested on two additional FR models with different losses or architectures, i.e., AdaFace [15] and MobileFaceNet [5]. All the models are pretrained on MS1Mv2 [7] dataset.

As comparison, five state-of-the-art explainable face verification methods, namely LIME [23], MinusPlus [20], xFace [16], CorrRISE [19], and xSSAB [14], have been launched and tested. In detail, the third-party adaptation from [20] for LIME is utilized and the official codes of MinusPlus, xFace, xSSAB, and CorrRISE have been adapted to compute in batch on GPU for acceleration purposes on multiple datasets.

B. Evaluation Methodology

The evaluation of the proposed explanation algorithm comprises two phases. Firstly, the visualization results of produced saliency maps are presented. In the second phase, the evaluation metric "Deletion&Insertion" proposed by [19] is employed for a quantitative comparison among state-of-the-art saliency-based XFV methods. It iteratively deletes/adds pixels from/to the inputs according to the saliency values and observes the impact on the overall verification accuracy. In general, the more precise the saliency map, the lower the "Deletion" score and the higher the "Insertion" score. However, gradient backpropagation-based methods produce more sparse salient points than perturbation-based methods and often obtain imbalanced "Deletion" and "Insertion"

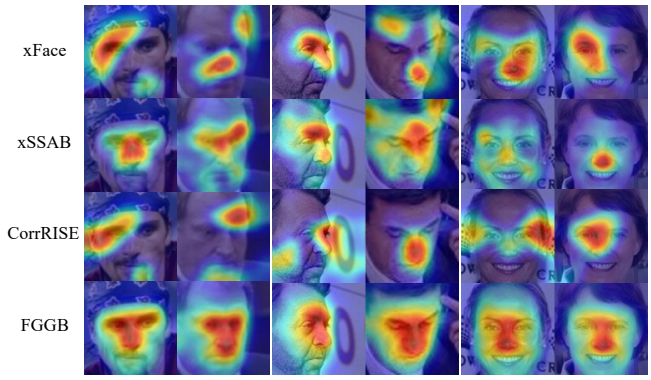

Fig. 3: Visual comparison of dissimilarity maps generated by FGGB and other XFV methods based on decisions of ArcFace model. Every two columns represent a pair of imposter faces. The saliency value increases from blue to red color.

TABLE I: Quantitative evaluation of similarity maps using Deletion and Insertion metrics (%) on LFW, CPLFW, and CALFW datasets. **Del (↓)** refers to the Deletion metric, the smaller the better. **Ins (↑)** refers to the Insertion metric, the larger the better. Red color denotes the highest score and blue color denotes the second highest score.

Methods	LFW		CPLFW		CALFW	
	Del	Ins	Del	Ins	Del	Ins
LIME [23]	35.82	80.76	27.61	71.46	34.21	78.40
MinusPlus [20]	29.64	83.28	24.63	69.27	29.06	79.57
xFace [16]	25.73	86.79	21.82	75.27	24.61	83.24
xSSAB [14]	25.98	84.49	23.00	72.20	26.09	80.25
CorrRISE [19]	24.51	86.80	20.01	77.07	24.30	83.23
FGGB	24.18	86.28	20.25	74.61	24.27	81.88

scores. Thus, we further improve the metric by applying a fixed-size Gaussian blur kernel to the end of all explanation methods before evaluation, guaranteeing a fair comparison.

C. Visual Demonstration

Given the same visualization tool, this section presents visualization results of saliency maps generated by the proposed FGGB method and other five XFV methods for face images randomly selected from CPLFW [31], LFW [13], Webface-Occ [12], and CALFW [32] datasets, representing different verification scenarios.

Fig. 2 shows similarity maps for genuine face pairs that the FR model "accepts". As a result, perturbation-based methods, such as MinusPlus, xFace, and CorrRISE, tend to generate too centralized similarity maps. Due to the fluctuating gradient issue of propagation-based methods, xSSAB exhibits some unnatural salient regions in the first and last examples. In comparison, FGGB provides clear contours for decision-critical facial regions and the produced saliency maps can always accurately highlight the most similar parts between two matching faces.

On the other hand, Fig. 3 presents dissimilarity maps for imposter face pairs that the FR model "rejects". MinusPlus and LIME are excluded because they do not provide dissimilarity maps. It is shown that dissimilarity regions highlighted by perturbation-based methods, CorrRISE and xFace, are generally spattered at different locations and are less intuitive for interpretation. On the contrary, FGGB can produce more stable and accurate dissimilar maps for non-matching face pairs, which is also supported by later quantitative evaluation.

TABLE II: Quantitative evaluation of dissimilarity maps using Deletion and Insertion metrics (%) on LFW, CPLFW, and CALFW datasets. **Del (↓)** refers to the Deletion metric, the smaller the better. **Ins (↑)** refers to the Insertion metric, the larger the better. Red color denotes the highest score and blue color denotes the second highest.

Methods	LFW		CPLFW		CALFW	
	Del	Ins	Del	Ins	Del	Ins
xFace [16]	75.53	92.64	53.18	87.42	64.26	90.91
xSSAB [14]	49.72	93.44	33.40	88.45	42.03	90.47
CorrRISE [19]	81.36	89.81	57.55	83.98	69.22	87.34
FGGB	44.03	93.35	28.71	88.88	34.55	90.06

TABLE III: Explainablity performance of FGGB tested on different face recognition models. The verification accuracy (%) of FR models and two explainability metrics (%) for FGGB are reported.

FR Models	Acc (LFW)	Deletion (↓)	Insertion (↑)
ArcFace [7]	99.70	24.18	86.28
AdaFace [15]	99.27	21.41	83.87
MobileFaceNet [5]	98.87	19.32	77.73

D. Quantitative Results

Table I quantitatively compares similarity maps obtained by FGGB and other XFV methods using the "Deletion&Insertion" assessment metric. The proposed FGGB method achieves superior scores in the Deletion metric on all three datasets while getting slightly lower Insertion scores when compared to CorrRISE and xFace. As for dissimilarity maps, Table II shows that FGGB provides the most accurate dissimilarity maps on multiple datasets and both metrics, which is consistent with the observations in the visual demonstration. Moreover, it is notable that FGGB outperforms another propagation-based method xSSAB in most scenarios, which proves the advancement of FGGB in addressing the noisy-gradient issue during backpropagation.

In addition, FGGB is further tested on two other face recognition models with different architecture and loss functions. Table III shows that when the FR model achieves similar verification performance, the saliency maps produced by FGGB also have similar explainability performance, which validates that FGGB is model-agnostic.

V. Conclusion

This paper contributes to the problem of explainable face verification by conceiving a new efficient and model-agnostic saliency explanation solution FGGB. It provides similarity and dissimilarity saliency maps to interpret both the "Accept" and "Reject" decisions made by the face verification system. Experiments show that FGGB exhibits excellent performance, particularly in dissimilarity maps, when compared to the current state-of-the-art. Moreover, this paper explores a new approach to mitigate the impact of fluctuating gradients during backpropagation, which provides insights for improving future gradient propagation-based explanation methods for general learning-based vision systems.

REFERENCES

[1] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. 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.

[2] N. Bousnina, J. Ascenso, P. L. Correia, and F. Pereira. A rise-based explainability method for genuine and impostor face verification. In *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6. IEEE, 2023.

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[4] A. Chattopadhay, A. Sarkar, P. Howlader, and V. 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.

[5] S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018.

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

[7] J. Deng, J. Guo, N. Xue, and S. 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.

[8] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.

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

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

[11] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[12] B. Huang, Z. Wang, G. Wang, K. Jiang, K. Zeng, Z. Han, X. Tian, and Y. 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.

[13] G. B. Huang, M. Mattar, T. Berg, and E. 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.

[14] M. Huber, A. T. Luu, P. Terhörst, and N. Damer. Efficient explainable face verification based on similarity score argument backpropagation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4736–4745, 2024.

[15] M. Kim, A. K. Jain, and X. 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.

[16] M. Knoche, T. Teepe, S. Hörmann, and G. 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.

[17] K. Li, Z. Wu, K.-C. Peng, J. Ernst, and Y. 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.

[18] Y.-S. Lin, Z.-Y. Liu, Y.-A. Chen, Y.-S. Wang, Y.-L. Chang, and W. 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.

[19] Y. Lu, Z. Xu, and T. 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.

[20] D. 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.

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

[22] V. Petsiuk, R. Jain, V. Manjunatha, V. I. Morariu, A. Mehra, V. Ordonez, and K. 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.

[23] M. T. Ribeiro, S. Singh, and C. 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.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. 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.

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

[26] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[27] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[28] J. R. Williford, B. B. May, and J. 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.

[29] Z. Xu, Y. Lu, and T. Ebrahimi. Discriminative deep feature visualization for explainable face recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2023.

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

[31] T. Zheng and W. 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.

[32] T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

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

[34] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.