XAIface: a framework and toolkit for explainable face recognition

Nelida Mirabet-Herranz*[1], Martin Winter[2], Yuhang Lu[3], Naima Bousnina[4], Jonas Pfister[5], Chiara Galdi[1], Jean-Luc Dugelay[1], Werner Bailer[2], Touradj Ebrahimi[3], Paulo Lobato Correira[4], Fernando Pereira[4], Felix Schmautzer[5], Erich Schweighofer[5].

[1]*Department of Digital Security*, *EURECOM*, Biot, France.

[2]*Institute for Digital Technologies, JOANNEUM RESEARCH Forschungs-GmbH*, Graz, Austria

[3]*Multimedia Signal Processing Group (MMSPG), EPFL*, Lausanne, Switzerland

[4]*Instituto de Telecomunicações – Instituto Superior Técnico*, Lisbon, Portugal

[5]*Centre for Computers and Law*, *University of Vienna*, Vienna, Austria

***Abstract*—Artificial intelligence-based face recognition solutions are becoming increasingly popular. Therefore, it is crucial to fully understand and explain how these technologies work in order to make them more effective and acceptable to society. This is the goal of the CHIST-ERA project XAIface, the final results of which are reported in this article: a framework and toolkit for improving AI decision explainability, in the context of automated face recognition, through several novel methods are presented. These methods are integrated into an end-to-end face recognition demonstrator system, which facilitates studying the impact of various influencing factors and system processes on recognition performance. By doing so, we can visually explain the decisions made by the face verification pipeline for specific instances in our test set using heatmaps and locally interpretable features. Furthermore, we offer a comprehensive explanation of the end-to-end model by examining the relationship between verification failures and misclassifications of soft biometric facial traits.**

***Index Terms*—Face recognition, Explainability, Soft biometrics, Human understanding, Visual explanation, XAIface**

I. Introduction

Face Recognition (FR) has become a key technology in our society, frequently used in multiple applications, ranging from access control and video surveillance to social media and automatic annotation, creating an impact in terms of privacy [1]. Automatically recognizing individuals in pictures and video raises several security and ethical issues, including but not limited to invasion of privacy and interference with national and European laws such as the GDPR [2] and the AI-Act [3]. A key and central issue in FR is that of trust. The degree of trust in FR solutions is still an open challenge, and several studies have shown that FR systems exhibit various weaknesses and biases toward specific genders, ethnicities, and ages of individuals [4]. As such, fairness and transparency must be tackled together with the development of better, more efficient, and accurate systems.

Moreover, the popularity of explainable artificial intelligence models (xAI) has grown in interest due to the need to explain the ever-complex artificial intelligence systems that have been developed and deployed by large companies or research groups [5]. The rapid improvements of these models correlate with their opaqueness. Current FR systems achieve outstanding performances [6] that can even exceed those of humans but are difficult to understand and analyze due to the opacity of the deep learning methods used [7]. Most state-of-the-art work regarding xAI is nowadays focused either on the development of transparent machine learning models or on the application of post-hoc explainability models. While transparency is most often an inherent property of simpler, classical machine-learning models, more complex state-of-the-art face recognition methods usually require post-hoc explainability techniques, such as saliency maps [8]. In addition, the analysis of the influencing factors relevant to the final decision of an AI-based face recognition system has been proven as an essential step to understand and improve the underlying processes involved [9], [10].

Following the above, in this paper, we aim to contribute to a better understanding of the decision mechanisms in face recognition based on deep learning in particular by proposing four different alternatives for explaining the output of our end-to-end face recognition pipeline based on the state-of-the-art FR network ArcFace [6]. In addition, we develop clear legal guidelines on the use and design of AI-based face recognition following the privacy-by-design approach. We believe that these insights will help increase the level of trust and social acceptance of FR technology.

The paper is organized as follows: in the first section, we provide motivation for our work and offer some general information regarding the topic of face recognition. In Section II, we present the related xAI models applied to FR and ethical and legal open issues in people recognition. Our FR framework, with special emphasis on the four different proposed explainability methodologies, is presented in Section III, while Section IV details the dataset used, protocol followed, and the implementation details. Results are presented and discussed in Section V, and finally, we conclude in Section VI with a summary of the outcomes of our study.

This article includes the outcomes of the XAIface European project and lists all the contributors to the project.

*Corresponding author: mirabet@eurecom.fr.

II. Related Works

A. xAI applied to FR

Explainable AI is an area of ongoing study, and several approaches have proven effective in addressing facial recognition problems. One of the earlier examples is Layer-wise Relevance Propagation (LRP), a framework that explains how deep neural networks make predictions for images by assigning relevance scores to image regions of a sample image. This approach has been applied to explainable facial expression recognition [11]. Other types of generic explainability methods focus on producing saliency map visualizations in the image space, such as deconvolutional networks [12]. Further notable approaches include e.g. Class Activation Maps (CAMs) [13], Grad-CAM++ [14], or Score-CAM [15], which provide improved visual explanations of CNN model predictions.

While previous methods required knowledge of the classifier's architecture, other types are model-agnostic and provide explanations without this information. One example of such a technique is LIME (Local Interpretable Model-Agnostic Explanations) [16], which analyzes the relationship between input data and predictions using a perturbation-based forward propagation approach. This method can be used with any model, acting as a surrogate for the original classifier. Also SHAP (SHapley Additive exPlanations) [17] is a prominent representative of this genus. It quantifies the positive or negative influence of each feature on a classifier's decision by focusing on measuring the relative impact of a given value against a baseline. This is related to Shapley-values in game theory. Another example of such a black box technique is Randomised Input Sampling for Explanation (RISE) [18], which generates saliency maps that indicate how important each pixel in an image is for the network's decision.

Although many explanatory approaches have been applied to face recognition (e.g., LIME and RISE), their applicability is limited because the elements used for explanation are not or only partially interpretable by humans. This issue has received only limited attention in the literature, with one notable exception being Williford et al. [19], who proposed a framework that generates saliency maps based on maximizing the match between a probe image, but coevally minimizing the match between the probe image and a gallery face image in which the region around a facial landmark has been in-painted to replace the original content.

B. Open issues in ethics and law

In the ethical domain of FR technologies it has long been established that explainability is central to building trust for users. A key aspect of explainability, aside from general transparency, is the ability to explain decisions to those directly or indirectly affected [20]. The legal framework of the European Union has long struggled to catch up with this ethical requirement. The first major milestone was the introduction of Art 22 GDPR in conjunction with Art 15 par 1 lit h GDPR, which grants data subjects the right to be informed about the existence of automated decision-making. If the scope of Art 22 par 1 GDPR is met, the controller must provide meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject. Art 22 GDPR has been subject to an extensive academic debate, starting with Goodman/Flaxman who were of the opinion, that the articles provide for a so-called "right to explanation", meaning a right to an ex-post explanation of the individual decision taken [21]. On the other hand, Wachter et al. stated, that no such right could be derived from the legal text [22]. While there was an argument to be made that the articles should be interpreted in such a manner, that an individual would have the necessary means to contest an automated decision [23], the opinion of Wachter et al. seemed to prevail. Furthermore, the utility of a right to explanation has been criticized. The Article 29-Working Party has highlighted the issue of providing complex explanations to individuals and urges controllers to focus on "clear and comprehensive" methods of delivering information to data subjects [24]. Hence, two central problems remained. Firstly, the content of a right to explanation must be defined and secondly, xAI in the FR technologies domain must deliver clear and comprehensive explanations in line with this definition.

The first issue was recently addressed by the judicial branch. The Advocate General (AG) of the European Court of Justice has elaborated in the SCHUFA-I-Opinion on the content of Art 22 GDPR in conjunction with Art 15 par 1 lit h GDPR. He established, that the information provided must "include sufficiently detailed explanations of the method used to calculate the [outcome] and the reasons for a certain result" [25]. The explanation must include the relevant factors and their respective aggregated weights in such a way, that the information is useful to challenge the decision. Hence, the AG derives an individual right to an ex-post explanation from the articles. It should be noted, however, that this issue has not been explicitly discussed in the judgment itself.

The method of addressing the second issue depends on the concrete system involved. Two common approaches usually highlighted in the legal xAI domain are saliency maps and counterfactual explanations [26] [23] [27]. For tasks like detection and recognition, employing saliency maps [28] as developed in this project may be one way to satisfy the requirements established by the AG. This is because they visually demonstrate relevant factors for the individual decision, show their respective aggregated weights with different color scheming and are interpretable by humans. Additionally, FR technologies users and developers must take into consideration future requirements of the AI-Act, namely the introduction of a new right to explanation according to Art 68c. Furthermore, new requirements on interpretability according to Art 13 and on accuracy according to Art 15 of the AI Act will drastically increase the need for xAI solutions.

III. Methodologies

A. End-to-End Explainable Face Recognition System

This section provides an overview of the end-to-end explainable face recognition system, which integrates various face

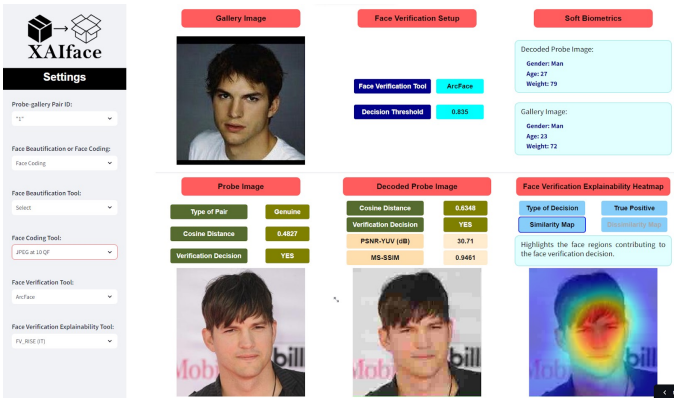


Fig. 1: Demonstrator user interactive interface layout

processing tools including face detection, pre-processing, and coding, with recognition and explainability modules.

First, the face image might undergo various processing steps, where the influencing factor analysis is carried out to understand their impact on the overall performance of the AI-based face recognition system. Then, the recognition module employs the state-of-the-art ArcFace model [6] with ResNet100 [29] feature extractor in the pipeline. This module performs standard face verification and identification tasks and its decision is subsequentially interpreted by the follow-up explainability modules. Four explainability modules are integrated into the end-to-end system providing comprehensive explanations for the face recognition model. These explainability techniques operate independently of each other and do not affect the face recognition performance. On the contrary, they provide various explanations from three distinct perspectives for the decision of the face recognition model and demonstrate powerful explainability capabilities.

Furthermore, this work designs and develops a demonstrator application for the proposed end-to-end explainable face recognition system, as depicted in Figure 1. This demonstrator application features a user interactive interface allowing users to perform face verification explainability experiments by selecting key elements such as probe-gallery pairs, face compression and beautification tools, and face verification and explainability tools. When the selection is made, it displays key conditions and results, including the face verification setup, selected face images and their verification results, soft biometrics characteristics, filtered or decoded images with verification and compression results, and explainability heat maps for the selected tool. This setup enables users to analyze and understand the performance and explainability of face verification processes. The developed demonstrator application is publicly available[1].

B. FV-RISE

Face Verification-based Randomized Input Sampling for Explanation (FV-RISE) [30] is a novel Face Verification (FV) explainability method to explain the decision-making process of any FV model without accessing or modifying the inner architecture of the model. The key novelty of FV-RISE is that it addresses both genuine and impostor FV attempts, as well as acceptance and rejection decisions using similarity and dissimilarity heat maps. The similarity and dissimilarity heat maps highlight the face regions contributing most to an acceptance and rejection FV decision, respectively.

The FV-RISE method is designed to estimate the pixels' importance for the FV decision by applying random masks to the probe image and measuring the impact of masking face regions on the FV model performance. The similarity and dissimilarity heat maps are generated through three key steps, notably *i)* Compute the reference similarity score between the original probe-gallery pair. *ii)* Generate a set of masked probe images by applying random masks to the probe image and compute the similarity scores for the new probe-gallery images. *iii)* Group the generated masks into masks for similarity and dissimilarity face regions based on their respective similarity scores compared to the reference score. Afterwards, perform the weighted sum for each group of masks to obtain the similarity and dissimilarity heat maps.

The decision-making is explained using a single heat map depending on the type of FV decision. Notably, the similarity heat map is used when a true/false acceptance decision is made, while the dissimilarity heat map is used when a true/false rejection decision is made.

C. CorrRISE

CorrRISE [31] is a new saliency map-based explanation method for face recognition, which provides similarity and dissimilarity saliency maps to interpret the decision of the deep face recognition system. While existing explanation methods for face recognition interpret the model's prediction with saliency maps indicating similar regions between any matching images, they are not often eligible for generating meaningful saliency maps for non-matching cases. The CorrRISE method aims to depict facial regions that the deep face recognition system deems similar or dissimilar between two given faces through the produced saliency maps. These maps can be used to analyze various scenarios, such as why the face recognition system believes two facial images are a good match or not, why it identifies matches even when faces are occluded or heavily compressed, and why it fails to give correct predictions in specific instances.

In principle, CorrRISE generates saliency maps by injecting perturbation on the input image and observing the impact on output. Thus, it provides "black-box" explanations and can be applied to any FR system without retraining or accessing 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's correlation coefficient between a list of similarity scores and random masks is calculated in a pixel-wise manner to obtain

[1]https://xaiface-demo.streamlit.app/

saliency maps. These saliency maps then disentangle similar and dissimilar pixels according to the correlation coefficients.

D. LIBF

Explainable Face Recognition by Locally Interpretable Boosted Features (LIBF) [32] is a method to obtain meaningful explanations and countermeasures by using a novel combination of locally interpretable, human-understandable and meaningful face-region descriptors (LIBF) and a lightweight and interpretable classification engine (EBM) [33].

In particular, LIBFs are a type of face features calculated around meaningful patch locations (e.g. eyes, cheeks, foreheads) in an image and projecting the local data to an embedding learned by a self-supervised technique. The first step is to use a state-of-the art face detector, such as RetinaFace [34], to extract and normalize the face image. This is followed by extracting face landmarks (e.g., eyes, nose, corners of the mouth) using the framework proposed by Bulat et al. [35]. Additional significant patches are then extracted based on a simple heuristic. The LIBFs themselves are then calculated by projecting these patches to embeddings learned through Grill et al. [36].

The matching process itself is straightforward by forming task-specific verification features for the EBM. In particular, since we need to provide both query and template LIBFs at the same time, we encode the comparison of the 187-dimensional LIBFs for query (Lq) and template features (L_t) by a simple concatenation in the verification feature representation used by EBM. This specialized representation can then be used to estimate the label (match or no match) encoded as $[0, 1]$ using the inherent explainability capabilities of the EBM. Please note, that more details of this approach can be found in our previous work and application paper [37].

E. Explaining via soft biometrics

The Explaining via Soft Biometrics (ESB) method aims to provide explanations by predicting and analyzing the performance differential of soft biometric trait estimation for different subgroups of the test set. This information is used to provide insights into whether there is a correlation between between face verification failures and misclassifications of soft biometric facial traits. This is achieved by computing the overall soft-biometric estimation performances across four categories (created from the results of the face verification pipeline): True Positive (FV-TP), True Negative (FV-TN), False Positive (FV-FP), and False Negative (FV-FN).

Unlike the visualization-based explanation techniques mentioned above, which provide local explanations about specific instances on the test set, the ESB approach targets a local but also a global explanation of the FR pipeline by searching a more general behavior of the model across the entire test set. This technique is employed after the FR pipeline, independent of the recognition process and other explainability methods.

IV. Experimental setup

A. Database

The images used in our experiments are selected from the publicly available IARPA Janus Benchmark-C (IJB-C) [38] dataset. It is a comprehensive and challenging dataset designed for the evaluation of face recognition algorithms. It includes images and videos with diverse conditions such as varying illumination, pose, and occlusions. The dataset comprises 3,531 subjects, 31,334 still images, and 117,542 video frames, making it significantly larger and more diverse than its predecessors. It is used to benchmark the robustness and accuracy of face recognition systems in a wide range of real-world scenarios, aiming to push the boundaries of face recognition technology.

B. Protocol

The various integrated explainability methods outlined in Section III operate based on the types of tasks and decisions predicted by the face recognition pipeline. To ensure a consistent explanation of the results and follow-up evaluation across different integrated explainability modules, we conduct the same face verification task and design a corresponding verification protocol for the entirety of the experiments. Specifically, the verification protocol comprises 3000 matching pairs and 3000 non-matching pairs of face images collected from the IJB-C dataset.

C. Implementation details

1) Implementation details for FV-RISE: In FV-RISE, the generated masks are primarily used to perturb random regions of the face on a pixel-wise basis, allowing for the measurement of their impact on FV performance. Consequently, the RISE [18] masking technique is employed to randomly generate these masks. To ensure precise similarity and dissimilarity heat maps, the number of generated masks is set to 10000. Additionally, these masks are initially created at a resolution of 5×5 before undergoing bilinear interpolation.

2) Implementation details for CorrRISE: The CorrRISE explanation method operates by injecting perturbations and does not require any training or access to the internal architecture of the face recognition model, but it relies on several key parameters to generate mask perturbations. This paper employs the default configuration of CorrRISE as stated in the original publication. The number of generated masks, i.e. the number of iterations, is set to 1000. For each mask, there are 10 patches and the size of each patch is 30×30 pixels.

3) Implementation details for LIBF: To capture more face parts beyond those provided by previous research [35], we selected six additional, relative patches (namely forehead, cheek, and chin - left and right, respectively) to be independent of resolution, scale, and rotation. We also identified which patches may be outside or invisible in extreme rotations and stored this information as an additional feature. Another important factor, namely the dimensionality of the embedding has been set to $d = 16$ as research shows, that 10 dimensions are sufficient for face recognition [39]. To efficiently communicate

the explanations to the end user, we developed a special visualization called Color Coded Face Part Contributions (CCFPC) explaining the magnitude of the contributions made by different parts of a person's face using intensity coding. Green and magenta is used to represent the support/no-support information for a distinctive decision.

4) Implementation details for ESB: For the soft biometric estimator, a popular open-source gender and age estimator is adopted: *DeepFace*[2]. For gender classification, *DeepFace* returns the labels "man" or "woman" along with their associated probabilities. For age estimation, an integer between 0 and 100 is returned once a human face is passed to the pre-trained models.

V. Experimental results

A. Heat map-based explanation results

The CorrRISE and FV-RISE explainability modules in the end-to-end explainable face recognition pipeline aim at providing heat map-based explanation results. This section illustrates a few examples of heat maps generated by the two modules for the verification process using the IJB-C dataset and protocol.

As shown in Figure 2a, the generated heat maps properly highlight the regions that the face recognition model perceived as similar between the matching pairs. In general, the similarity regions concentrate on nose, beard, eyes, and glasses, while there are distinct variations from image to image. For example, the explanation heat maps show that the face recognition model believes the beard region of the second pair of images is similar enough to be classified as the same person. As for non-matching examples in Figure 2b, the explainability module is capable of clearly highlighting the most dissimilar regions in their faces, such as the nose and mouth of the second example, and the mouth and jaws of the third example. Moreover, the heat map-based explainability module of the explainable face recognition system is also robust to various image qualities, such as ill-posed images in the first and third pairs of Figure 2a and low-resolution images in the third pair of Figure 2b.

B. Results on LIBF explanation

As already stated and justified in our previous work [37], LIBF explanations are difficult to apply directly on verification tasks if the embedding method used by LIBF is based on a different database and data space than the one being verified. However, in setups requiring very high distinctiveness, setting the normalized recognition threshold for explainable matches very strictly to 0 or 1 for non-matching and matching faces respectively allows for feasible explanations even though more 'ambiguous' samples in between might be missing. In other words, LIBF explanations are most useful for explaining high-confidence decisions of other non-explainable but high-quality verification methods.

To communicate the explanation clearly, we utilize the importance of individual features (patches) obtained from the EBM. Thus we create color-coded overlays of the face-patch regions and assign green/red color intensities based on the contributions of the 10 most important individual features to the decision. Green shaded patches indicate support for the correct decision made by the classifier, while magenta colored patches indicate support for an incorrect decision.

[2]https://github.com/serengil/deepface

TABLE I: Accuracy (in %) of the gender estimation and Mean Absolute Error (in years) of the age estimation for the different subgroups considered. *FV* = face verification.

GENDER				**AGE**			
FV Match		**FV Non-Match**		**FV Match**		**FV Non-Match**	
79.11		81.10		12.68		12.57	
FV-TP	**FV-FN**	**FV-FP**	**FV-TN**	**FV-TP**	**FV-FN**	**FV-FP**	**FV-TN**
80.92	73.72	74.47	80.85	13.17	11.00	11.14	12.51

Figure 3 provides two illustrative examples of positive and negative matches, allowing the user to e.g. easily see, that the forehead and left corner of the mouth in the first example of Figure 3, contribute most to the correct matching decision while the eyes show some differences voting for dissimilarity. In contrast, the right forehead and eyes correctly vote for the dissimilarity of the two faces, but the left cheek would erroneously indicate, that the two faces are the same.

C. ESB study

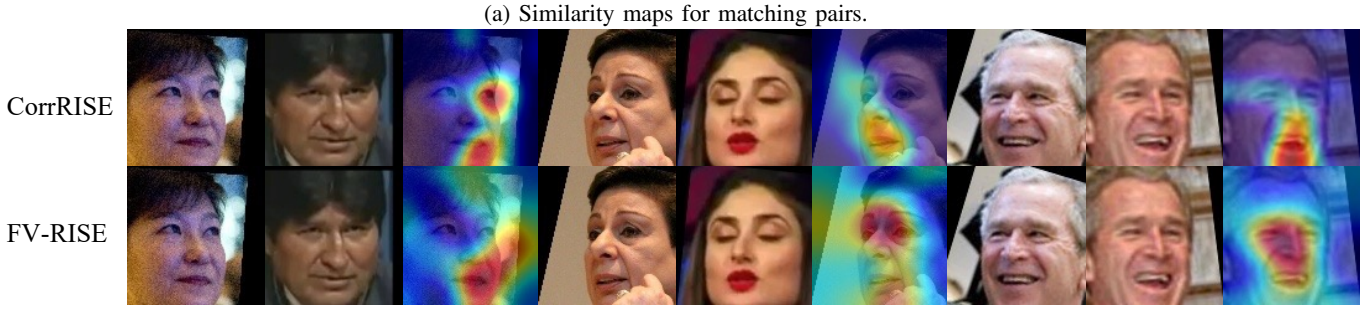
In Figure 4, we present the explanation from the ESB model for various pairs of faces, both matching and non-matching. The estimated gender and age are indicated below each image.

In Table I, we report the accuracy of the gender predictions and the Mean Absolute Error (MAE) in years of the age estimator for each class considered. We can observe how the accuracy and MAE for the matched set of images and the non-matched set of faces are very similar for the gender and age predictions, respectively. However, when we consider each of the four subsets, we observe that in the case of correct face verification output, i.e., TP and TN, the gender accuracy is higher, while when the face verification fails (FP, FN), the probability of gender misclassification is higher. This indicates a possible correlation between the difficulty in verifying identity and the gender for challenging face images. However, from Table I, we see how this behavior is opposite for the age estimation method. FN and FP exhibit a lower error in years when the age is estimated compared to correctly verified images. We can conclude that the behavior of gender estimation correlates more strongly with the accuracy of the face verification task compared to age estimations.

VI. Conclusion

This article presents the results of the XAIface project aimed at enhancing the transparency and interpretability of face recognition systems. We engaged in a discussion summarizing the open ethical and legal issues that have emerged during our project, emphasizing the importance of addressing these concerns to ensure responsible deployment of face recognition systems. In addition, an end-to-end explainable FR demonstrator is presented, integrating the developed xAI models to enable users to interact with them ensuring understanding and

(a) Similarity maps for matching pairs.



(b) Dissimilarity maps for non-matching pairs.

Fig. 2: Heat map explanations of CorrRISE and FV-RISE results for both matching and non-matching face pairs are presented in a standard face verification scenario. Each pair of probe and gallery images and the corresponding heat map for the probe image are represented by every three columns in the figure. The similarity and dissimilarity heat maps explain why the verification model makes correct predictions on matching and non-matching faces by highlighting the more similar or dissimilar aspects for the model in the probe-gallery pair.

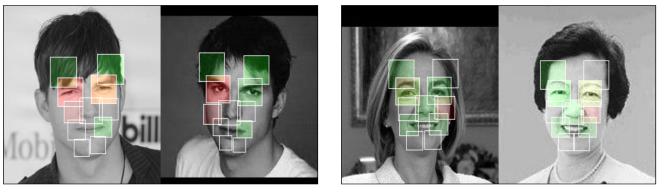


Fig. 3: LIBF explanations for matching and not-matching face pairs. In the case of identical faces (left), patches supporting successful verification are colored green, while non-supporting ones are red. The opposite is true for different faces (right).

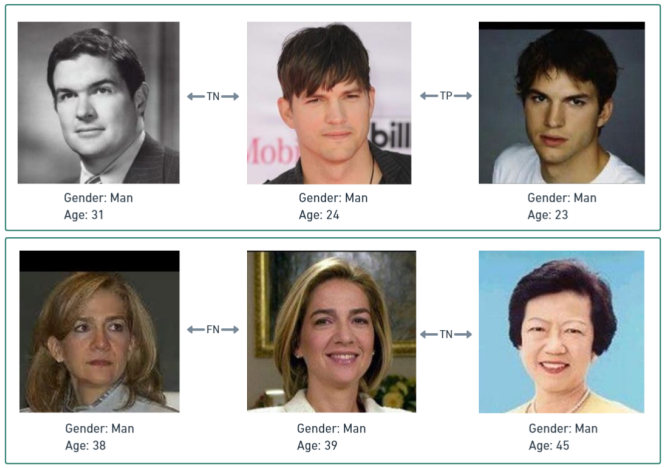


Fig. 4: Explaination from the ESB model for different matching and non-matching face pairs. The estimated gender and age are presented below each image. The output of the face verification is depicted between each pair in between arrows.

trust in the FR pipeline. We shared the results of experiments demonstrating the feasibility of our approach, illustrating the effectiveness of individual modules in enhancing the interpretability and performance of face recognition systems. Through these efforts, we aim to advance the field of xAI and promote responsible deployment of face recognition technology. Future perspectives include the unification and fusion of different visual explanations, as well as the possibility of combining local explanations such as heatmaps with global explainability techniques like soft biometric estimation. More information about the project can be found on the project website: https://xaiface.eurecom.fr/.

ACKNOWLEDGMENT

This work has been partially supported by the European CHIST-ERA program (grant agreement CHIST-ERA-19-XAI-011). Some other parts of the work have been supported by European Union´s Horizon 2020 research and innovation program under grant number 951911 - AI4Media and the Swiss National Science Foundation (SNSF) 20CH21 195532. The industrial supporters of the XAIface project are IDEMIA, HENSOLDT, IN groupe SURYS and Quantum Integrity

REFERENCES

[1] C. Galdi, P. Lobato Correia, Y. Lu, J. Pfister, and M. Winter, "Xaiface: Measuring and improving explainability for ai-based face recognition," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) OJ L 119, 4.5.2016, p. 1–88.

[3] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final) (hereafter 'AI-Act').

[4] P. Grother, M. Ngan, and K. Hanaoka, *Face recognition vendor test (fvrt): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD, 2019.

[5] P. C. Neto, T. Gonçalves, J. R. Pinto, W. Silva, A. F. Sequeira, A. Ross, and J. S. Cardoso, "Explainable biometrics in the age of deep learning," *arXiv preprint arXiv:2208.09500*, 2022.

[6] 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*, 2019.

[7] M. Huber, M. Fang, F. Boutros, and N. Damer, "Are explainability tools gender biased? a case study on face presentation attack detection," in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 945–949.

[8] Y. Lu and T. Ebrahimi, "Explanation of face recognition via saliency maps," in *Applications of Digital Image Processing XLVI*, vol. 12674. SPIE, 2023, pp. 218–229.

[9] N. Bousnina, J. Ascenso, P. L. Correia, and F. Pereira, "Impact of conventional and ai-based image coding on ai-based face recognition performance," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[10] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay, "Impact of digital face beautification in biometrics," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2022, pp. 1–6.

[11] F. Arbabzadah, G. Montavon, K.-R. Müller, and W. Samek, "Identifying individual facial expressions by deconstructing a neural network," in *German Conference on Pattern Recognition*. Springer, 2016.

[12] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

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

[14] 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)*. IEEE, 2018, pp. 839–847.

[15] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[16] 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*, 2016, pp. 1135–1144.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, 2017.

[18] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018.

[19] J. Williford, B. May, and J. Byrne, "Explainable face recognition," in *Proceedings ECCV*. Springer, 2020, pp. 248–263.

[20] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2018.

[21] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine, Vol 38, No 3*, 2017.

[22] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *International Data Privacy Law, Vol 7, No 2*, 2017.

[23] B. Gyevnar, N. Ferguson, and B. Schafer, "Bridging the transparency gap: What can explainable ai learn from the ai act?" in *European Conference on Artificial Intelligence (ECAI)*, 2023.

[24] Art-29-Working-Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01), 2018.

[25] Opinion of the AG of the ECJ, 16 March 2023, C-634/21 "SCHUFA-I".

[26] P. Hacker, "Comments on the final trilogue version of the ai act," *Europeannewsschool.eu*, 2024.

[27] H. Asghari, N. Birner, A. Burchardt, D. Dicks, J. Faßbender, N. Feldhus, F. Hewett, V. Hofmann, M. C. Kettemann, W. Schulz, J. Simon, J. Stolberg-Larsen, and T. Züger, *What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making*, 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6375784

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

[29] 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*, 2016, pp. 770–778.

[30] B. Naima, A. João, C. Paulo Lobato, and P. Fernando, "A rise-based explainability method for genuine and impostor face verification," in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2023, pp. 1–6.

[31] 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*, 2024, pp. 4726–4735.

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

[33] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "Interpretml: A unified framework for machine learning interpretability," *arXiv preprint arXiv:1909.09223*, 2019.

[34] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27. [Online]. Available: https://doi.org/10.1109/ASYU50717.2020.9259802

[35] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

[36] J. B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised Learning," 2020, to do. [Online]. Available: http://arxiv.org/abs/2006.07733

[37] H. Neuschmied and M. Winter, "Explainable face verification for video archive documentation," in *20th International Conference on Content-based Multimedia Indexing*. Orleans, France: Association for Computing Machinery (ACM), Sep. 2023.

[38] M. Brianna, A. Jocelyn, D. James A., K. Nathan, M. Tim, O. Charles, J. Anil K., N. W. Tyler, A. Janet, C. Jordan, and G. Patrick, "IARPA janus benchmark-C: Face dataset and protocol," in *International Conference on Biometrics*, 2018, p. 158–165.

[39] D. Celis and M. Rao, "Learning facial recognition biases through VAE latent representations," in *1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*, 2019, pp. 26–32.