A Deep Adversarial Framework for Visually Explainable Periocular Recognition

Anonymous CVPR 2021 submission

Paper ID ****

Abstract

The ability to portray the reasoning behind a decision has been at the core of major research efforts. It serves not only to increase trust amongst the stakeholders in AI, but to potentially improve the entire system as a whole. In this work, we present our efforts towards explainable periocular recognition, with a simple, yet performant framework that automatically provides transparent outputs. Quantitative and qualitative results are shown to validate the proposed goals, which reiterate the notion that explainability should be strongly considered when designing ML algorithms.

1. Introduction

This work focuses on developing an integrated framework for recognition and subsequent explanation. In this context, one should consider the system's accuracy, as well as, its ability to portray the reasoning that supports a match/non-match decision. The latter is becoming an integral part of Machine Learning systems, given how ubiquitous and dependable they have become in recent years [2]. Thus, we diverge from the black-box paradigm and embrace an explainable nature, as seen in Fig. 1.

Typically, a recognition task involves a set of unique and non-transferable features that, when given to a system designed to do so, can unmistakably identify a subject. Biometrics, as they are designated in the field, serve such purposes, as long as they are universal, sufficiently distinguishable, resilient to changes and realistically easy to collect [1]. Upon proving compliant with the aforementioned requirements, biometrics can be separated into two dominant categories. On one hand, *physiological* features like the irises, fingerprints and retinas are naturally possessed by a given subject. On the other hand, the gait and signature typify *behavioural* biometrics, largely manifested when a subject interacts with the surrounding environment [18].

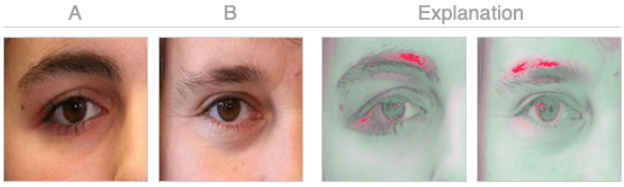

Figure 1. Visual explanation provided by the proposed method. Areas marked with red tones provide reasons as to why the two samples come from different subjects (i.e. eyebrows, eyelashes and a skin spot).

As a subordinate field, periocular recognition makes use of the rich area comprising and surrounding the eye, in which the iris, sclera, eyebrow, eyelid and skin stand out. We also acknowledge this set of facial components, considering, where applicable, their colour and/or shape.

Regarding explainability and its applicability in recognition tasks, let us not forget that Deep Learning solutions rely on model complexity and abstraction prowess to become truly accurate. Although seemingly innocuous, there could be seriously negative outcomes if opaque algorithms gamble on the clearance of unauthorised people into sensible areas. It becomes clear, then, that including explainable components into AI systems is imperative. More recently, the politicians have addressed this urgency in formal terms. Namely, the EU, through the GDPR [5], introduced the notion of "right to an explanation". The definition and scope are still subject to debate [20], but these are definite strides to regulate the explainable depth of autonomous systems.

Following the above paragraphs, this paper describes a simple framework that receives a pair of images (either *genuine* or *impostor*, depending on whether the images come from the same subject or not) and produces a twofold output: a binary match/non-match decision and a visual explanation (in this case, it takes the form of an image). This can be seen as the main contribution of the present work, in that it creates an accurate and explainable solution. Other important insights include the use of the powerful genera-

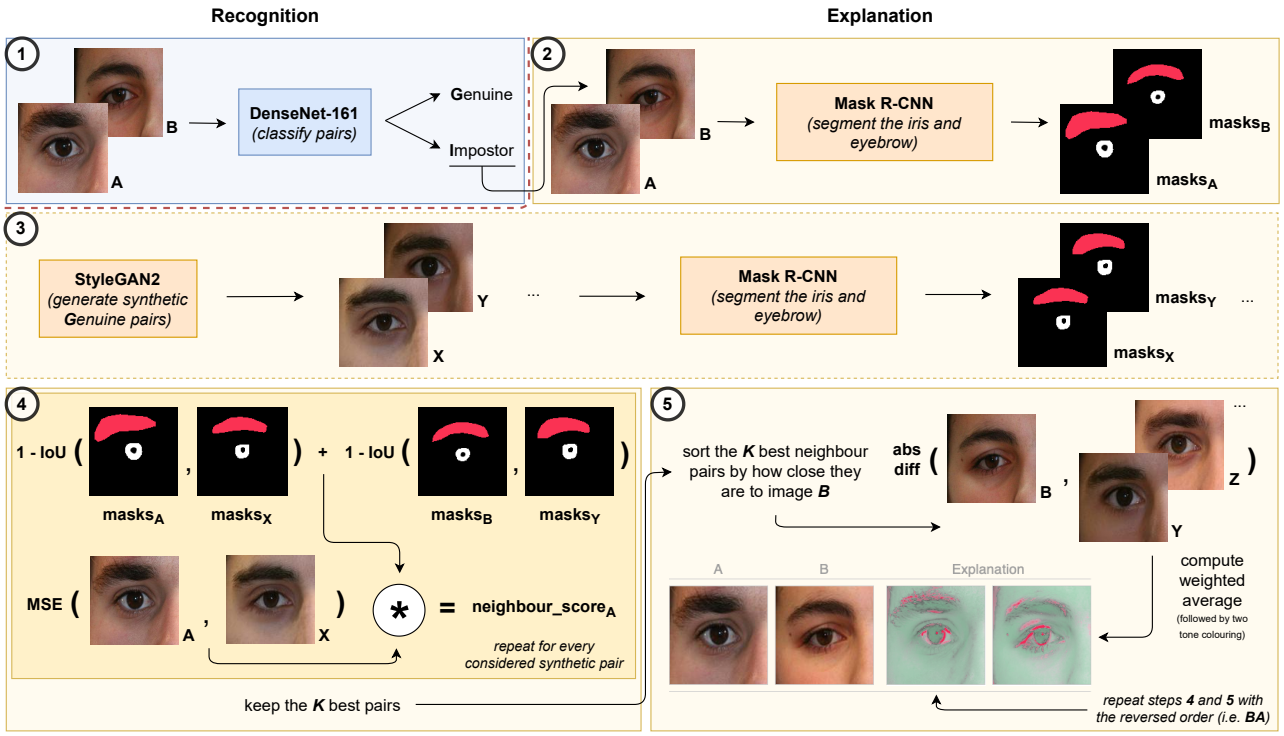

Figure 2. Diagram of the main pipeline. Step one (i.e. recognition) encompasses a CNN that distinguishes between *genuine* and *impostor* pairs. Then, if deemed *impostor*, steps two to five (i.e. explanation) try to find *genuine* synthetic pairs that closely resemble the test pair. By doing so, and despite looking similar, the test pair will probably contain certain internal differences (i.e. between images A and B) that the synthetic ones do not, thus providing an interpretable explanation.

tive capabilities that GANs possess to create samples that wouldn't otherwise exist in the training data, thus giving it more variety and flexibility (such process can be seen as a form of data augmentation).

Fig. 2 illustrates the main steps that enable the periocular recognition task and posterior explanation: a CNN is trained to optimally produce a match/non-match decision and, if the pair is deemed to be *impostor*, a search process finds similar looking *genuine* pairs. The key here is that, if the test pair has significant disparities between images A and B, the closest synthetic pairs most likely do not (they are *genuine*). Then, by simply computing pixel-wise differences between the test pair and the closest neighbours, those disparities become evident. This assumption is justified by the CNN's non-match decision, which can likely be attributed to differences in iris colouration, presence/absence of skin spots or eyebrow thickness.

The remainder of this paper is organised as follows: section 2 summarises the research efforts in the fields of periocular recognition and ML Explainability; section 3 describes our method; section 4 analyses the main results and section 5 closes this paper with some final remarks.

2. Related work

2.1. Periocular recognition

The seminal breakthroughs in periocular recognition tasks can be traced to a set of methods termed *feature descriptors*. Methods such as HoG, LBP and SIFT were able to produce simplified representations by relying on edges, textures and keypoints, respectively. In [22], the results from each feature descriptor are fused to provide a more comprehensive description of an image's content. This work served as the basis for consequent fusion based approaches, as in [6]. Extending towards the field of supervised learning, in [11] a Restricted Boltzmann Machine is used to learn a probabilistic distribution over the input data, further discriminated with metric learning and SVMs.

With the realistic applicability of Deep Learning schemes, researchers turned to popular architectures, like CNNs, in the pursuit of ever increasing recognition accuracy. Accordingly, in [23] the main concept involves the use of multiple CNNs that are trained to become specialised in classifying certain semantic information (e.g. gender, age, and more). Then, a score fusion process creates a unified architecture. In [17], the authors enforce a CNN to ignore the ocular region (due to this region's tendency to capture

light reflections and other performance degradation factors) and rely more on the surrounding area (eyebrow, eyelid and skin). Going against the idea of only relying on either the ocular or periocular regions, in [19] the iris and periocular biometrics are separately explored for classification purposes, with the resulting scores being fused to reach a final decision. More recently, in [4] the authors bridge the gap between biometric recognition (in their case, facial) and interpretability, by learning feature specific filters that activate in a range of preferred spatial locations, and, in [9], an integrated solution is proposed by leveraging part discovery as a form of attention. However, a fusion of periocular recognition and interpretability remains largely unexplored, thus motivating the development of novel solutions.

2.2. Machine learning explainability

In the literature, explainable techniques are commonly discriminated in terms of depth, scope and model applicability [12], [15]. Depth is related to the length to which we explain a given model, i.e. whether the technique limits the model's complexity to make it more transparent (*intrinsic*) or allows complexity and focuses on explaining just the outputs (*post hoc*). Scope indicates the range that a technique possesses, i.e. if it explains individual predictions (*local*) or the model's entire behaviour (*global*). Finally, the applicability metric classifies techniques based on their model affinity, i.e. whether they are only compatible with a specific family of models (*model-specific*) or virtually any kind of model (*model-agnostic*). Commonly cited techniques include LIME [14] and SHAP [13]. The former uses a surrogate linear model, trained on perturbed data (e.g. disabled clusters of adjacent pixels), to locally approximate the behaviour of a complex black-box model. The latter uses game theory and Shapley values, which are assigned to the features based on how important they are to a given prediction. Additionally, Saliency Maps [10] use the derivative of a highly complex function (essentially, a CNN) with respect to a given input image, to determine which pixels need to be changed the least, while also changing the output class the most. Finally, for plotting reasons and, therefore, outside the preferred scope of this work, PDP [7] and ALE [3] are able to produce plots correlating independent variables and a dependent, target variable, by exploiting the notions of marginal and conditional distributions, respectively.

3. Proposed method

3.1. Learning phase

The main learnable components of the proposed method comprise the DenseNet-161, Mask R-CNN and StyleGAN2 models: the CNN is trained to, essentially, perform a verification task, the segmentation model is optimised to produce high-quality masks for the iris and eyebrow and, finally, the GAN learns how to create synthetic data that, while inspired by the training set, is diverse enough to approximate unseen subjects. Additionally, a fourth, auxiliary model (i.e. ResNet-18) is fitted to discriminate between images from the left and right sides of the face. Although trained separately, all four models learn from the same training split, which excludes a small set of IDs that are reserved for performance evaluation tasks.

3.2. Inference

Once trained, the framework evolves into five major steps, as depicted in Fig. 2. Firstly, the DenseNet-161 model is used to verify the claimed identity: upon receiving a pair of images, the model's output is one of two classes (*genuine* or *impostor*). If the pair is deemed to be *impostor*, the remaining steps are responsible for creating a visually interpretable explanation.

According to the numbering system used, step two takes the test pair and, using Mask R-CNN, segments the irises and eyebrows. Then, step three uses the StyleGAN2 generator to create a large, synthetic dataset of *genuine* pairs (i.e. where both images have differences but belong to the same person). For each of these synthetic pairs, the ResNet-18 model determines its side configuration (i.e. whether both images come from the left or right side of the face) and, as before, masks are obtained with the segmentation model. Note that it is highly advised to perform this step beforehand and not during the inference stage (for speed reasons).

After obtaining the synthetic images and their respective masks, the synthetic dataset is structured based on the iris positions, enabling faster search. To that end, the clustering algorithm K-Means is trained on a subset of the iris segmentation masks to compute three centroids, one for each major iris position (i.e. left, centre and right). With them, one can store the images based on their combination of iris positions (e.g. left-left, right-centre, etc...). By doing so, when searching, the algorithm can just rely on the synthetic pairs that share the same combination as the test pair, saving time and useless calculations. Obviously, the centroids provide a liberal classification of where an iris is. During the search process, the irises are subject to a stricter comparison.

Upon settling for a portion of the synthetic dataset that closely meets the iris position constraint, the segmentation masks can be further used to determine which generated pairs have the iris and eyebrow in, approximately, the same position as their counterparts in the test pair. Such precondition is key, given that pixel differences, which make up the visual explanation, are sensitive to component misalignment. With effect, a synthetic neighbour's score is given by the following formula:

$$\text{score} = w_{\text{masks}} * \text{MSE}(\text{test\_pair}_{\text{A}}, \text{neighbour}_{\text{X}}) \quad (1)$$

In equation 1, it becomes clear that a weighted distance

is measured for each synthetic neighbour, with respect to the first images (i.e. image A for the test pair and image X for the synthetic one). Moreover, w_{masks} influences the computed MSE score to favour pairs that have good component alignment and penalise those that do not (to achieve this, a factor equal to $1 - \text{IoU}(.,.)$ is used). Then, if the score is better than the currently saved pairs, the considered pair is kept as one of the best matches, up to that point. This iterative process continues until every considered synthetic pair is analysed. In practice, the search process tries to find, amongst the thousands of synthetic pairs, the ones that are closest to the test pair, in terms of the first images. Therefore, given that the second image of the test pair is not a *genuine* match, it will most likely be different in some areas to those of the optimal synthetic neighbours, and that is exactly the kind of dissimilarities that make up the final explanation.

From this point forward, the K closest neighbours are sorted by how close they are to image B, using equation 1 (computing the MSE score with respect to the second images, instead). Finally, to produce the final explanation, the K best neighbours are used to compute pixel differences against the test pair's image B. In practice, a neighbour's distance is subtracted from the total sum of distances, thus creating an inverted distance. This approach ensures that neighbours with smaller distances receive more weight as opposed to those with bigger distances. Then, the inverted distances are simply divided by the sum of inverted distances so as to normalise them. The final difference image results from those inverted distances acting as weights to determine the importance of each intermediate difference. To add visual appeal, the resulting explanations are coated with red and green tones, thus making them even more understandable.

3.3. Implementation details

The DenseNet-161 model is trained for 15 epochs with a learning rate of 0.0002 and a batch size of 64 image pairs. The Adam optimiser is responsible for the weight optimisation process (with default β_1 and β_2 values). A similar training setup is used to train the ResNet-18 model, albeit for a smaller number of epochs (i.e. 5). The Mask R-CNN's training process uses almost all the default values, translating into a learning rate of 0.001, a batch size of 1 and 30 epochs worth of training (in this case, fine-tuning from the COCO pre-trained weights). As for the StyleGAN2 architecture, its training comprises a total of 80000 iterations and a batch size of 8. After converging, the generator is capable of synthesising realistic looking images, such as the roughly 400000 pairs that make up the artificial dataset. Finally, the number K, which determines how many synthetic neighbours should be kept, receives a default value of 15.

4. Experiments and discussion

4.1. Datasets

As mentioned above, the proposed framework consists of two modules, one for recognition and the other for explanation purposes. Regarding the former, the chosen CNN is solely trained on the UBIPr dataset [16], which is naturally oriented towards periocular recognition problems and contains valuable ID information. As for the latter, it mainly relies on a combination of UBIPr and FFHQ [21]. Despite not being immediately applicable to the context of this work (i.e. it contains full face images, thus requiring extra steps to extract the periocular region), the FFHQ dataset contains unquestionable variety in terms of attributes, some of which are scarcer in the UBIPr dataset. In practice, a small, but curated, portion of the FFHQ samples is used to create a more varied super set (Fig. 3). Regardless of their source, all images are resized to a common shape, depending on the task (i.e. 512x512x3 for Mask R-CNN, 256x256x3 for StyleGAN2 and 128x128x3 for the CNNs).

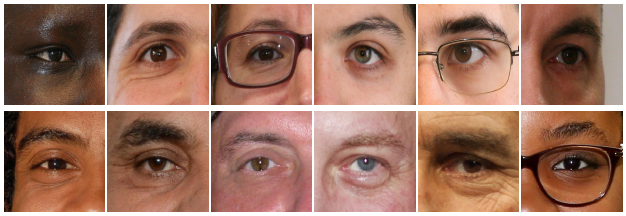

Figure 3. Samples from both datasets used. The top row represents the UBIPr dataset, whereas the bottom row comes from a cropped version of the FFHQ dataset.

4.2. Working scenario

As is customary with biometric recognition systems, it is important to define the working mode and world setting, upon which the system is built. With respect to the working mode, a system is said to be in verification mode (also referred to as *one-to-one*) if it tries to validate a claimed identity (i.e. subjects identify themselves and the system's task is to validate those claims by comparing the extracted features to those stored in a database). On the other hand, if in identification mode (*one-to-many*), a system will try to retrieve the most likely match, within a know set of possible IDs [1]. As for the world setting, it discriminates between systems that can only accept a pre-defined set of IDs (closed-world) and those that continue to function even when new, unexpected subjects appear (open-world).

Based on the definitions above, our method essentially performs verification (the comparisons are done *one-to-one*) within an open-world setting, meaning that unseen subjects do not impose compromises or limitations.

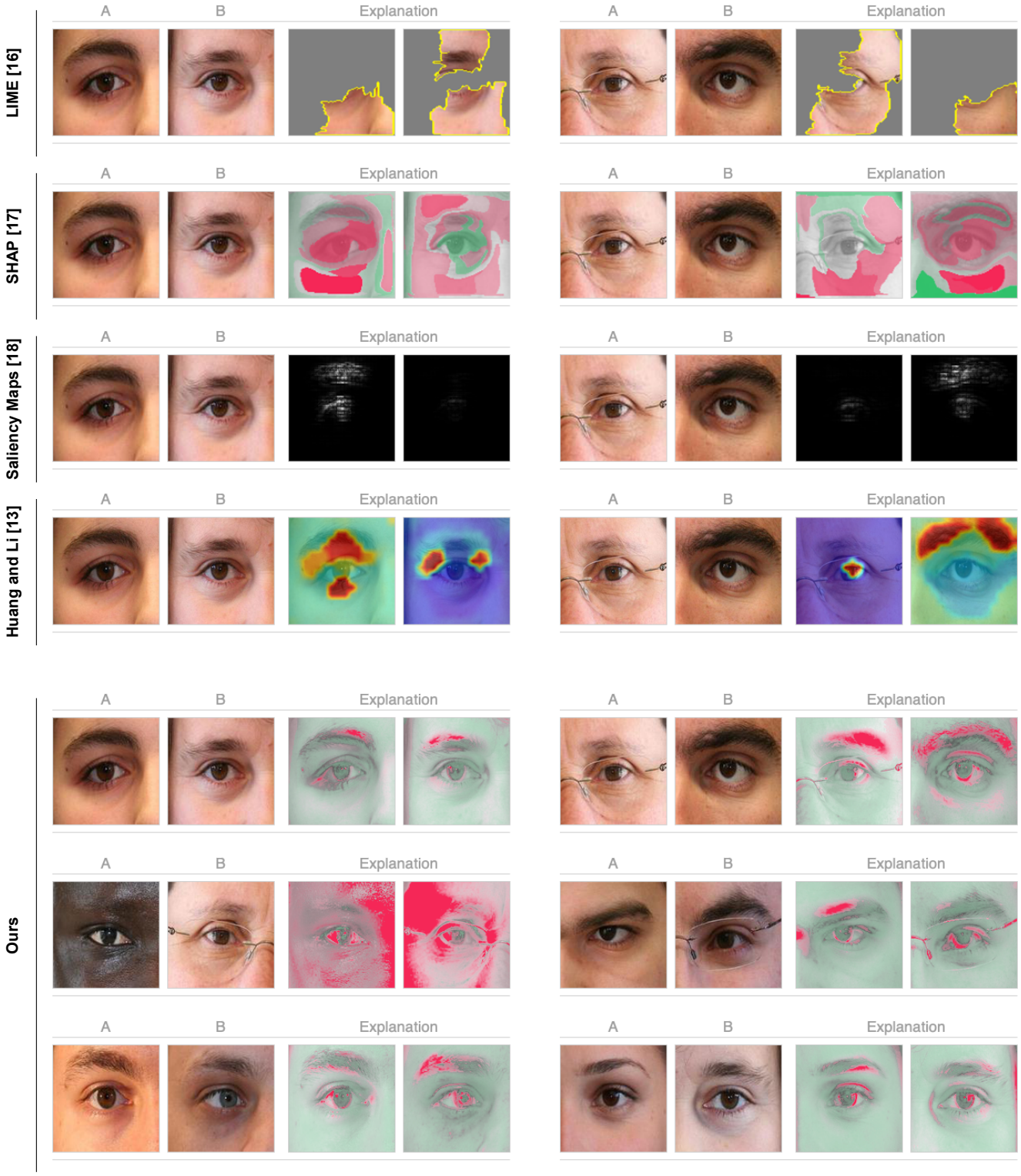

Figure 4. Results from three standard interpretability techniques (i.e. LIME, SHAP and Saliency Maps), a state-of-the-art interpretable deep model for fine-grained visual recognition (i.e. [9]) and our method. Notice how our results are really clear in highlighting components that justify a non-match decision (e.g. skins, eyebrows, irises, skin spots and, even, eyelashes).

4.3. Qualitative evaluation

Like most explainable techniques, the proposed method produces explanations whose accuracy relies on some degree of subjectivity. Nonetheless, a DenseNet-121 model, trained to perform the verification task, is paired with either LIME, SHAP or Saliency Maps to create a comprehensive comparison scheme, to which we add the method described in [9].

Fig. 4 displays the expected results from a visually explainable system. In practice, LIME tries to keep the most important super-pixels, SHAP highlights those it deems important in red tones and Saliency Maps produce greyscale explanations. As for the method by Huang and Li, it generates a heat-map in which red tones elevate important areas. Focusing on the common pairs between all methods, the left sample is essentially different with regards to eyebrow thickness and presence/absence of a noticeable skin spot. As for the right one, the most obvious disparities have to do with the eyebrow areas. Overall, our results are the most informative, when compared with the remaining four solutions. While LIME and SHAP do a decent job, Saliency Maps provide a faint explanation. It is Huang and Li's method that comes closer to our level of visual appeal, by clearly highlighting portions of the eyebrow and a portion of subject A's skin spot, in the left pair. Moreover, when given the right sample, it generates a solid red area comprising subject B's eyebrow. However, upon closer inspection, our results show more appealing visual cues: in the left sample, distinct red tones on top of A's skin spot and eyelashes, as well as, reiterated eyebrow differences in the right sample with highlights in both eyebrows, rather than just one. As for the remaining samples, the third (just below the first) is clearly explained by highlighting the entirety of both skin areas, which are obviously different between images A and B. Finally, in the fourth pair it is also shown how the eyelids differ, by colouring that periocular component on subject B's image, and, in the fifth sample, subjects B's eyebrow and iris are accurately shown in red.

4.4. Quantitative evaluation

A comparison between the proposed method and an existing periocular recognition alternative is summarised in Table 1. A bootstrapping methodology was employed by sampling 90% of the available dataset and diving the resulting samples between two subsets: training (80%) and test (20%). Then, the CNN was trained as usual and its EER and AUC scores were saved. Such process was repeated 10 times, culminating in means and standard deviations for both metrics. Note that, for these experiments, our method was trained using the UBIRIS.v2 dataset [8], following the same scheme as the other method. Analysing the obtained results, one can conclude that the recognition module of our framework surpasses its competitor with regards to the EER metric, even when considering an open-world setting (which the other method does not). It should be noted that, due to a modular design, the recognition module in the proposed framework can be replaced to achieve superior performance, while maintaining the explainability properties intact.

Method	EER	AUC
Ours (open-world)	$0.108 \pm 3\text{e}{-2}$	$0.813 \pm 5\text{e}{-2}$
Ours (closed-world)	$\mathbf{0.087 \pm 2\text{e}{-2}}$	$\mathbf{0.910 \pm 2\text{e}{-2}}$
Zhao and Kumar [23]	$0.109 \pm 2\text{e}{-3}$	–

Table 1. Comparison between the proposed method (in both world settings) and a state-of-the-art strategy (strictly operating in a closed-world setting).

4.5. Ablation study

The two major hiper-parameters of the proposed method are K and the length of the synthetic dataset. Changes to these values can affect the quality of the generated explanations in a less than optimal way (as seen in Fig. 5).

4.5.1 Number of closest neighbours

The value K determines how many synthetic pairs should be considered the closest *genuine* neighbours to a test pair classified as being *impostor*. Overall, smaller values lead to more sensitive and jagged results, unlike those achievable with larger values. Up to a certain point (e.g. 15), increasing K creates smoother explanations, due to the larger number of samples taken into account when averaging the intermediate differences. This trend, however, starts returning incremental improvements (or none at all), that do not justify themselves (notice how, in Fig. 5, the output with K set to 50 or more stops presenting a prominent tone on the eyelid).

4.5.2 Synthetic dataset length

Considering the main goal of the search process is to find *genuine* pairs that closely resemble a test pair, restricting the amount of possible matches can impose certain limitations. With effect, in Fig. 5, it becomes clear how working with a smaller set of possible matches leads to less evident highlights, especially around the eyelid and, at a smaller scale, the eyebrow. Therefore, due to the use of relatively light calculations (i.e. MSE and IoU), the increased search times, that accompany larger datasets, do not make this an unfeasible solution, while also allowing for better results.

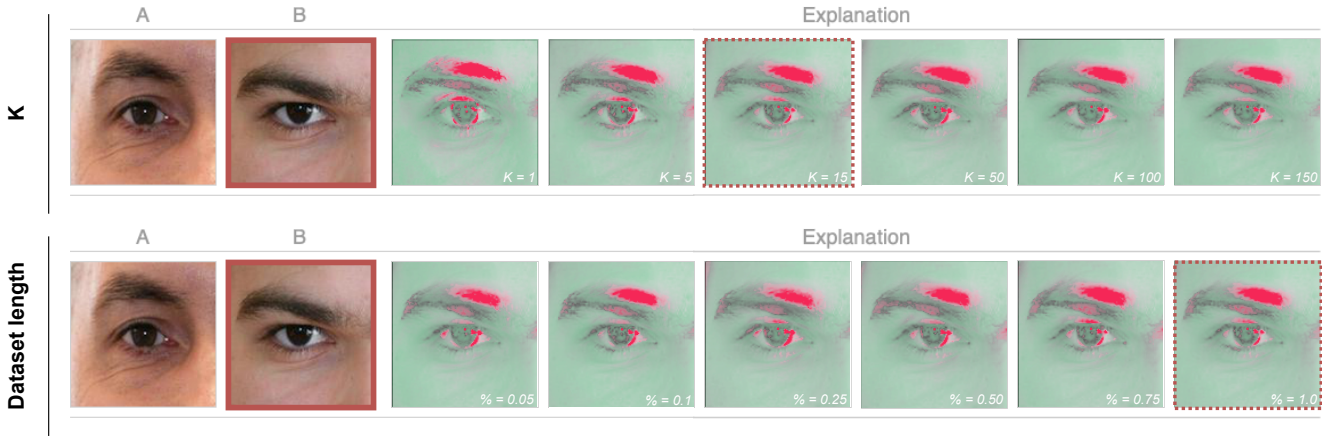

Figure 5. Visual perturbations when key aspects of the proposed method are changed. The red square indicates which image is being explained (i.e. B), while the red dashed square indicates the default values. In general, increasing K up to 15 allows for smoother explanations, as does keeping a relatively large dataset. Reducing the latter tends to produce less decisive results.

5. Conclusions and further work

In this paper, a method for explaining periocular recognition decisions is proposed. By harnessing the generative power of GANs to create synthetic pairs that come from the same subject but do not possess crucial differences (like those present in *impostor* pairs), we are able to generate clear explanations. Additionally, the modular nature of the proposed method ensures that, if required, the periocular recognition stage can be replaced by other designs without compromising the explanations. Furthermore, a similar approach to the one used in this work can be explored for related tasks, like face recognition.

References

[1] A. Ross A. K. Jain and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004. 1, 4

[2] D. Braines R. Tomsett A. Preece, D. Harborne and S. Chakraborty. Stakeholders in explainable ai. arXiv:1810.00184 [cs], Set. 2018. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1810.00184. 1

[3] D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models. arXiv:1612.08468 [stat], Ago. 2019. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1612.08468. 3

[4] H. Li X. Shen B. Yin, L. Tran and X. Liu. Towards interpretable face recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9347–9356, 2019. 3

[5] European Commission. General data protection regulation. 2018. Accessed on: Feb. 27, 2021. [Online]. Available: https://gdpr-info.eu. 1

[6] A. Ross et al. Matching highly non-ideal ocular images: An information fusion approach. *2012 5th IAPR International Conference on Biometrics (ICB)*, pages 446–453, 2012. 2

[7] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. 3

[8] R. Santos J. Oliveira H. Proença, S. Filipe and L. A. Alexandre. The ubiris.v2: A database of visible wavelength iris images captured on- the-move and at-a-distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1529–1535, 2010. 6

[9] Z. Huang and Y. Li. Interpretable and accurate fine-grained recognition via region grouping. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8659–8669, 2020. 3, 5, 6

[10] A. Vedaldi K. Simonyan and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 [cs], Abr. 2014. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1312.6034. 3

[11] A. Kumar L. Nie and S. Zhan. Periocular recognition using unsupervised convolutional rbm feature learning. *2014 22nd International Conference on Pattern Recognition*, pages 399–404, 2014. 2

[12] Z. C. Lipton. The mythos of model interpretability. arXiv:1606.03490 [cs, stat], Mar. 2017. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1606.03490. 3

[13] S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. Long Beach, California, USA. 3

[14] S. Singh M. T. Ribeiro and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. arXiv:1602.04938 [cs, stat], Ago. 2016.

Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1602.04938. 3

[15] C. Molnar. Interpretable machine learning. a guide for making black box models explainable. 2019. Accessed on: Feb. 27, 2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/. 3

[16] C. Padole and H. Proença. Periocular recognition: Analysis of performance degradation factors. *Proceedings of the Fifth IAPR/IEEE International Conference on Biometrics – ICB 2012*, 2012. New Delhi, India. 4

[17] H. Proença and J. C. Neves. Deep-prwis: Periocular recognition without the iris and sclera using deep learning frameworks. *IEEE Transactions on Information Forensics and Security*, 13(4):888–896, 2018. 2

[18] H. Su M. Bennamoun S. Minaee, A. Abdolrashidi and D. Zhang. Biometrics recognition using deep learning: A survey. arXiv:1912.00271 [cs], Fev. 2021. Accessed on: Feb. 27, 2021. [Online]. Available: http://arxiv.org/abs/1912.00271. 1

[19] B. C. Dhara R. K. Rout S. Umer, A. Sardar and H. M. Pandey. Person identification using fusion of iris and periocular deep features. *Neural Networks*, 122:407–419, 2020. 3

[20] B. Mittelstadt S. Wachter 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*, 7(2):76–99, 2017. 1

[21] S. Laine T. Karras and T. Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. Long Beach, CA, USA. 4

[22] A. Ross U. Park and A. K. Jain. Periocular biometrics in the visible spectrum: A feasibility study. *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6, 2009. 2

[23] Z. Zhao and A. Kumar. Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 12(5):1017–1030, 2017. 2, 6