Efficient Explainable Face Verification based on Similarity Score Argument Backpropagation

Marco Huber[1,2], Anh Thi Luu[1], Philipp Terhörst[3], Naser Damer[1,2]
[1] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[2] Department of Computer Science, TU Darmstadt, Darmstadt, Germany
[3] Paderborn University, Paderborn, Germany
Email: marco.huber@igd.fraunhofer.de

Abstract

Explainable Face Recognition is gaining growing attention as the use of the technology is gaining ground in security-critical applications. Understanding why two face images are matched or not matched by a given face recognition system is important to operators, users, and developers to increase trust, accountability, develop better systems, and highlight unfair behavior. In this work, we propose a similarity score argument backpropagation (xSSAB) approach that supports or opposes the face-matching decision to visualize spatial maps that indicate similar and dissimilar areas as interpreted by the underlying FR model. Furthermore, we present Patch-LFW, a new explainable face verification benchmark that enables along with a novel evaluation protocol, the first quantitative evaluation of the validity of similarity and dissimilarity maps in explainable face recognition approaches. We compare our efficient approach to state-of-the-art approaches demonstrating a superior trade-off between efficiency and performance. The code as well as the proposed Patch-LFW is publicly available at: https://github.com/marcohuber/xSSAB.

1. Introduction

Automated face recognition (FR) has become an increasingly important part of our lives. It can be used to unlock a smartphone, cross borders at automated border checkpoints, or pay with a face. This is due, among other things, to the ease of use and high accuracy of modern FR systems. In recent years, the high accuracy of biometric systems has been driven primarily by larger databases [2], innovative solutions [4, 27], and advances in deep learning [5, 9, 13, 24].

However, the methods based on deep learning have the disadvantage of being difficult to understand because they include millions of parameters and are highly complex models [31]. For various reasons, biometric systems require that they are more understandable to humans. Understanding increases trust, can highlight unfair or unequal behavior toward different subgroups, or help develop better systems [1, 3].

Figure 1. An example visualization of our proposed efficient explainable face verification explanation, xSSAB, on the novel Patch-LFW benchmark and its evaluation protocol. The added patches from an imposter identity on an original genuine pair shifted the model's decision to non-match. During the evaluation, the added imposter clues (patches) are replaced by the original pixels based on the identified dissimilar area (pink) of our approach.

The reasons or clues why two face images are determined to be a match or non-match by an FR system have received increasing attention recently [17, 23, 25, 29, 30]. Especially in security-relevant areas, but also for usability, it is interesting to understand why two images are falsely recognized as a match or falsely recognized as a non-match (e.g. why didn't I match my passport at the automatic border control gate?).

In the field of biometrics, explainability and interpretability of biometric decisions and systems have been identified as outstanding problems [20]. In the area of more understandable face matching decisions, Lin et al. [25] proposed xCos. Their approach is based on a novel more interpretable cosine metric that provides meaningful explanations. Huber et al. [17] have proposed modeling the uncertainty and confidence of model decisions using stochastic forward passes to gain more insight into the decision process. In comparison, Knoche et al. [23] proposed a confidence score and a visualization approach based on systematic occlusions. A current trend is to follow a black-box approach [23, 29, 30], where only the input is changed and the changes in the output are observed, without a deeper understanding of the inner workings of the model. While these

approaches provide meaningful explanations, they require plenty of forward calculations which are time-consuming. Different approaches are white-box explanation approaches [25], which require access to the model, but can be more efficient. Our proposed approach follows the white-box approach to provide fast explanations without additional training.

One remaining challenge in the field of explainable FR is having an evaluation protocol to assess the validity of the explanations. Often, they are evaluated only on the basis of visualizations on a subjective basis without a quantitative evaluation [23, 25, 29, 30], which makes it difficult to compare different approaches objectively.

In this work, we 1) propose a novel, efficient, and training-free white-box approach based on similarity score backpropagation to indicate areas in a pair of face images to explain, which parts of the image are interpreted as similar or dissimilar for an FR model, with an example shown in Figure 1, 2) propose an objective evaluation dataset and an evaluation protocol that allows comparing the performance of different explainable face matching approaches based on visualization.

Our contributions are:

- 1) A training-free explainable FR approach that is competitive with state-of-the-art (SOTA) on several FR models while being time-efficient
- 2) The first explainable FR benchmark, Patch-LFW, including an evaluation protocol that allows to compare explainable face matching approaches in an objective manner.

2. Related Work

In recent years, the performance of FR systems has been improving and has already surpassed the performance of humans [26]. This improvement is based on larger data sets, larger models, and new loss functions, such as ArcFace loss [8], CurricularFace loss [15], MagFace loss [27] or ElasticFace loss [4]. With the increase in verification performance, the models became less understandable due to their complexity and size, which raises questions about the inner workings and causes of decisions.

The first direction of research that focused on making the face matching process more understandable were works related to uncertainty mapping. The idea for uncertainty mapping comes from thinking that faces may be ambiguous or lack identity information. Moreover, the problem of model and data uncertainty is of general interest in the deep learning community [12, 21]. Shi and Jain [36] proposed probabilistic face embeddings, where each face image is represented as a Gaussian distribution with the feature as the mean and the uncertainty of the features as the variance. This general idea has then been adopted in several ways [6, 7], including propagating the uncertainty present in the process of face matching to the comparison itself [17, 18]. Others tried to explain the performance variations in FR over different demographic groups [11], or assign the different in explainability performs [16] as well as the different levels of face image quality [10, 38].

The second research direction towards explainable FR followed the trend in computer vision to visualize important areas using saliency or heatmaps [35, 40], there are also works that visualize the crucial areas in the decision-making of FR. Applying methods known from computer vision like GradCAM [35] or Score-CAM [40] are not natively applicable for FR models in an optimal manner, since they are designed for classification problems and not for a process that consists of feature (embedding) extraction followed by embedding comparison (matching), which is typically applied in SOTA FR systems.

To visualize the important areas for an FR model, Lin et al. [25] proposed a novel similarity metric named xCos based on a learnable module to provide meaningful explanations. Their approach can be applied to most of the verification models, however, the module has to be trained, which reduces its adaptability, if the underlying model is often changed. Knoche et al. [23], Mery [29] and Mery and Morris [30] proposed to explain black-box FR models, following a justification explainability approach [33] designed to visually communicate the decisions' evidence instead of an introspective approach reflecting the inner workings visually [33]. Their approaches are designed to work without access to the FR model and are based on perturbing or altering the input face images and investigating changes in the output. While this produces interpretable saliency maps, the calculation is time-consuming. This poses a problem in practical applications, for example, if the user has to wait a long period of time for a reaction from the system, or if the data has to be transferred to a more powerful system in order to be processed there.

All the saliency map-producing approaches mentioned above did not evaluate their approaches in a quantified way, but rather limited their evaluation to visualizations, making an assignment and comparison in terms of the correctness and quality of the proposed approaches hard. They also did not present or use a well-defined benchmark specifically designed to demonstrate their results in a comparable manner.

3. Methodology

In this section, we present and rationalize our proposed xSSAB approach to explain face verification decisions efficiently. Understanding which parts of an image pair lead to a matching or non-matching decision is crucial to make FR more transparent and explainable. We propose to backpropagate the similarity score of face image pairs through a Siamese FR setup to efficiently indicate which parts of the image pair contribute to a match or non-match decision. By investigating the positive and negative impact of feature dimensions on the similarity-based comparison score, gradi-

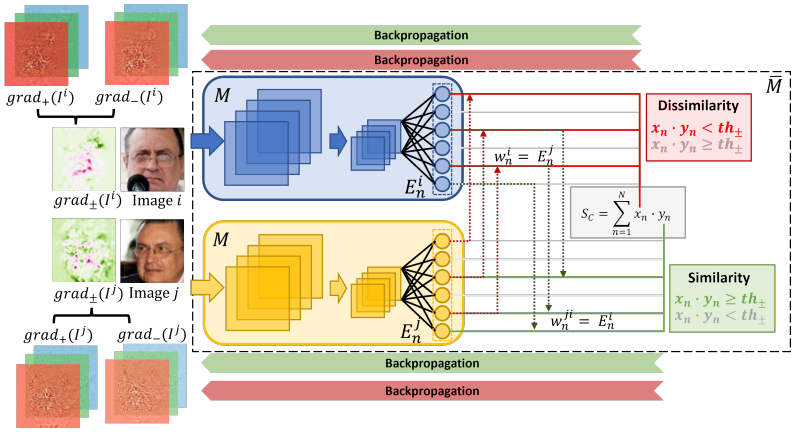


Figure 2. Overview of the proposed approach: In a siamese fashion, both face images are processed by a FR system M that is extended with an additional cosine similarity layer, which simply calculates the cosine similarity between the face embeddings.. Based on the system decision threshold and the contribution of the different features to the similarity score, gradients are backpropagated to obtain maps that highlight similarity and dissimilarity. These maps are then fused in a final step to get a single map, highlighting both.

ents based on these impacts can be backpropagated to highlight important pixels. An illustration of our methodology is shown in Figure 2.

In the first step, we propagate both images through a siamese network of FR model M to gain feature embeddings E^i, E^j of the face images I^i, I^j as demonstrated in Figure 2. We extend the FR model M by adding a cosine similarity layer, which calculates and outputs the cosine similarity S_C of both input face images to include the matching process in the network architecture. To simplify the calculation we also internally normalize the embedding vectors, E^i and E^j, before calculating the cosine similarity which simplifies the cosine similarity formula:

$$S_C(E^i, E^j) = \frac{E^i \cdot E^j}{\|E^i\|\|E^j\|}, \tag{1}$$

to

$$S_C(E^i, E^j) = \sum_{n=1}^{N} E_n^i \cdot E_n^j, \tag{2}$$

as $\|E^i\|\|E^j\| = 1$. N denotes the number of dimensions of the embeddings, E^i and E^j. E_n^i and E_n^j denotes the feature value at position n of E^i or E^j, respectively. Our extended FR model $\overline{M}$ can therefore be defined as:

$$\overline{M}_{i,j} = \sum M(I^i) \cdot M(I^j), \tag{3}$$

where $M(I^i)$ is similar to the embedding E^i and $M(I^j)$ to the embedding E^j:

$$\overline{M}_{i,j} = \sum_{n=1}^{N} E_n^i \cdot E_n^j. \tag{4}$$

Since $\overline{M}_{i,j}$ includes weights dependent on the provided input images to calculate the cosine similarity ($\overline{M}_{i,j}$ output) of the provided input images, I^i and I^j, it is dependent on the input.

Given the simplified cosine similarity function (Equation 2), we can observe that, expectedly, the cosine similarity will increase if the feature values x_n and y_n share the same direction and, otherwise, decrease the score. Since the final decision of match or non-match is not just dependent on the similarity score but also includes the system-dependent decision threshold th_d, we include this in the calculation of our approach. With the assumption, that each feature in the different feature dimension ideally at least contributes equally to the matching decision and therefore to the similarity score, we define an argument threshold $th_\pm$. This $th_\pm$ is set to:

$$th_\pm = \frac{th_d}{n}, \tag{5}$$

where n refers to the number of dimensions in the feature space. We then define a feature argument a_n for each feature dimension n as:

$$a_n = E_n^i \cdot E_n^j. \tag{6}$$

The interpretation is then that the feature argument a_n in the feature dimension n with a positive impact on the score ($a_n \geq th_\pm$) provides a *positive argument* and with a negative impact on the score ($a_n \leq th_\pm$) as provides a *negative argument*. The intuition is, that the argument threshold $th_\pm$ defines the minimum strength of an argument a_n to be considered a positive argument ($a_n \in a_+$) depending on the system-dependent threshold th_d. This is the case if the argument is at least as strong as it has to be if all other arguments equally contribute to a matching decision. In the other case, it is considered a negative argument ($a_n \in a_-$).

Starting from the calculated positive argument set a_+ and negative argument set a_-, we calculate the gradient based on the arguments, either only for the set of positive arguments (a_+) or for the set of negative arguments (a_-). We calculate the gradients backward through the FR model $\overline{M}_{i,j}$ to obtain the pixels that have the most influence [37] on the positive or negative arguments given both images I_i and I_j. To limit the calculation to only the impact on the

positive or negative arguments, we manipulate the weight w^i of the cosine layer of $\overline{M}_{i,j}$ (which is originally equal to E^j, since it computes S_C), so only the weights are included in the forward and backpropagation process that are either positive or negative:

$$w_n^i = \begin{cases} E_n^j & \text{if } a_n \in a_+ \\ 0, & \text{otherwise} \end{cases}, \tag{7}$$

if we want to obtain the similarity map of image I^i given image I^j and

$$w_n^i = \begin{cases} E_n^j & \text{if } a_n \in a_- \\ 0, & \text{otherwise} \end{cases}, \tag{8}$$

if we want to obtain the dissimilarity map of image I^i given image I^j. We define the models with the adjusted weights w_n^i as $\overline{M}_{i,j}^+$ in the first case and $\overline{M}_{i,j}^-$ in the second case. Since we manipulate the weights depending on the given comparison image, we get two different models and maps for the images I^i and I^j.

To calculate the gradient based on the positive ($grad_+$) and negative arguments ($grad_-$), we compute:

$$grad_+(I^i) = \frac{\partial \overline{M}_{i,j}^+}{\partial I^i}, \tag{9}$$

for the similarity map and

$$grad_-(I^i) = \frac{\partial \overline{M}_{i,j}^-}{\partial I^i}, \tag{10}$$

for the dissimilarity map. This allows us to calculate gradient-based maps only based on the features that have a positive or negative impact regarding a matching or non-matching decision based on the system's decision threshold.

To optimize the visualization further, we take the mean of the gradients of the three color channels (c). To be independent of the sign of the gradients, we also take the absolute value:

$$\overline{grad}_+(I^i) = \frac{1}{3}\sum_{c=1}^{3} |grad_+, c(I^i)|, \tag{11}$$

for the positive explanation map $\overline{grad}_+$. The negative (non-matching) map, $\overline{grad}_-$, is calculated in the same manner, only based on $grad_-, c(I^i)$. Our approach provides us with two explanation maps, showing which pixels lead to a positive or negative argument regarding the final matching decision of the FR system depending on the given decision threshold th_d and the investigated face images I^i and I^j. To get a single explanation map, $\overline{grad}_\pm$, explaining both, similarities and dissimilarities, we combine the two calculated explanations maps naturally by subtracting the negative explanation map from the positive explanation map:

$$\overline{grad}_\pm = \overline{grad}_+ - \overline{grad}_-. \tag{12}$$

To have a less fragmented visualization, we finally apply a Gaussian blur filter with 5×5 filter and with $\sigma = 5$, similar to Knoche et al. [23].

The proposed xSSAB approach, therefore, generates a single face verification explanation map based on the pixels' influence on the final matching or non-matching decision based on the system-dependent decision threshold. To achieve this, gradients based on the similarity score arguments are utilized. Calculating the explanation maps rather based on the internal behavior than just interpreting the model in a black-box fashion and altering the input, allows more efficient transparency of the models' behavior.

4. Experimental Setup

4.1. Face Recognition Models

To show the validity and the generalizability of our approach, we utilize four different SOTA FR models in our experiments. All models share the same ResNet-100 [13] architecture and have been trained with the corresponding loss functions. All the models were trained on the MS1M-V2 [8] dataset. The used models are: ArcFace [8], ElasticFace-Cos [4], ElasticFace-Arc [4], and CurricularFace [15], and they are all used as pre-trained models provided in their respective official repositories. We utilize these models to show that our approach can be applied to a wide range of diverse FR models without training or fine-tuning and the models are competitive to other SOTA models such as MagFace [28] or AdaFace [22].

4.2. Evaluation Benchmark: Patch-LFW

For the evaluation of the proposed explainable face matching approach and the comparison with other methods, we build a new benchmark dataset, *Patch-LFW*. Patch-LFW is based on the Labeled Faces in the Wild (LFW) [14] dataset. Since in most cases it is more interesting, from a practical perspective, to know why the system made a wrong decision than to understand why it made the right decision, we artificially increase the errors that the system makes. The reason we chose LFW as the starting point for our Patch-LFW is the simplicity of the dataset. To measure how well we explain wrong decision reasons, we need to minimize unknown reasons for wrong decisions and only have the ones we control, thus we need a benchmark that SOTA FR makes as minimum as possible wrong decisions. SOTA systems solve the verification problem in the LFW dataset almost perfectly, which minimizes other influences apart from the patches we add, which is what we need. Choosing a more challenging dataset as the baseline would make it harder to distinguish if the approaches can identify the added deterioration (patches) or inherent "clues" such as bias [39] or occlusions [32, 41]. To synthetically increase the amount of "false non-matches", in each of the genuine pairs, patches from a random image of a different identity have been added to the reference image. To synthetically increase the amount of "false matches", in each of the imposter pairs, patches from the same image have been added to the reference image. In total, we randomly added 27

(a) Genuine Pairs (b) Imposter Pairs

Figure 3. Example pairs from Patch-LFW: the images with the yellow outline in LFW are in Patch-LFW replaced by the images with the red outline. By adding information from a random imposter identity (genuine case) or from the same image (imposter case) the amount of false matches (FM) and false non-matches (FNM) increases.

patches of size 16×16 pixels which may overlap per image. Before adding the patches, the images of LFW [14] are preprocessed following the procedure of Deng et al. [8]. Examples of the newly created images and the original images from LFW [14] are shown in Figure 3. Given the new Patch-LFW, the "false match" rate as well as the "false non-match" rate at a fixed threshold determined at the Equal Error rate (EER) [19] on using the whole LFW dataset [14] increased drastically as shown in Table 1, which is our goal.

4.3. Evaluation Protocol: Decision-based Patch Replacement

To quantitatively determine the quality of explanation maps, we evaluate on our newly proposed Patch-LFW with a novel evaluation method (Decision-based Patch Replacement (DPR) curve). The DPR curves are inspired by the Insertion and Deletion curves [34]. First, the similarity maps for the original image pairs in the Patch-LFW dataset are determined. Based on these maps and the decision threshold th_{EER} determined on the original LFW without patches for each FR model, we proceed as follows:

- 1) If the patched image pair is considered a match based on th_{EER}, the most similar pixels are adjusted based on the explanation map. If the patched image pair is considered a non-match, the least similar pixels are adjusted. During the adjustment, the identified pixels are replaced by the original pixels from LFW.

- 2) After adjusting a fixed amount of pixels (in our case 5%), the FR systems are utilized to evaluate the new performance in terms of FMR and FNMR. If the explainability approach is of high quality, it should be able to detect the areas (added patches) that lead to "false matches" and "false non-matches" and reduce the errors when more and more pixels are replaced. If a pixel is detected to be replaced that is original and not part of an added patch, it is left unchanged. This step is repeated until all pixels have been replaced.

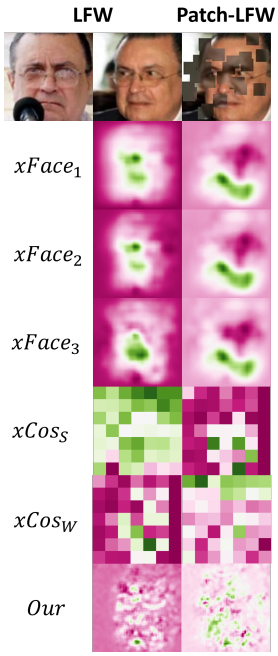


Figure 4. Comparison of different approaches: explanation maps visualized for different approaches. Pink areas indicate dissimilarity, and green areas indicate similarity. The values are not normalized. The design of the maps is very different. While the $xFace_x$ approaches tend to highlight contiguous areas, the highlighting in $xCos_x$ is patch-wise. Our approach, on the other hand, is more fine-grained. More examples are provided in the supplementary material.

To determine the quality, we now look at the drop in error rates over the proportion of pixels replaced. If the curve is lower, it indicates a better performance of the explanation map as the error rate is lower. This proves that the explanation map detected the artificially inserted patches and replaced them with the original pixels, which removes the clues that led to the artificially created error. Such plots are presented in Figure 5 and will be discussed in Section 5.

4.4. Explainable Face Verification Methods

To evaluate our proposed efficient explainable face verification approach based on similarity score argument backpropagation (xSSAB), we compare our approach in terms of quality and latency with two SOTA solutions: xCos [25] and xFace [23].

xCos [25] modifies the backbone of the model with a 1×1 convolution to preserve the position information of each feature. With an additional attention mechanism, two outputs are retrieved: an attention weight map W and a patched cosine map S. Since the authors only provided a pre-trained version for an ArcFace model [8], we stick to the comparison with this model. We experimentally tested to use the proposed pre-trained cosine module and metric on the other models as well, but the equal error rate on

	LFW	Patch-LFW	
Model	EER	FMR	FNMR
ElasticFace-Arc [4]	0.003	0.434	0.649
ElasticFace-Cos [4]	0.002	0.639	0.475
CurricularFace [15]	0.005	0.643	0.444
ArcFace [8]	0.005	0.660	0.606

Table 1. Performance of the utilized models on LFW [14] and the proposed benchmark Patch-LFW. The FMR and FNMR at Patch-LFW have been evaluated at the EER threshold th_{EER} determined on LFW. By definition, the FMR and FNMR of LFW are equal to the EER. The results show that the introduced patches increased the FMR and FNMR on all models, providing an experimental setup for explainable face matching methods.

ElasticFace-Arc [4] was 0.2533, which is not competitive when compared to the performances presented later in Section 5. We investigated the performance by inserting pixels either based on the weight map ($xCos_W$) or based on the patches cosine map ($xCos_S$). Because both of the maps provide equally-sized patches that share the same value, we do not iteratively replace 5% of the pixels but replace them patch-wise. An example of both maps is shown in Figure 4.

xFace [23] follows the black-box approach and considers the FR models as pure input-output functions. The core principle of their proposed approach is based on occluding parts of the image and investigating the deviation of the occluded and non-occluded images. In total, they proposed three variations of their approach, which we refer to as $xFace_1$, $xFace_2$, and $xFace_3$. For the parameters of their approach, we follow their proposed values and select stride $s = 5$ and patch sizes $p \in \{7, 14, 28\}$. An example of the visualization of the maps can be seen in Figure 4.

We also compare our approach against a random approach (RND). The random approach randomly selects pixels to replace and serves as a minimum baseline.

4.5. Evaluation of Latency

To evaluate the efficiency, we perform the computations on the same machine equipped with an NVIDIA Quadro P4000 GPU and an Intel Core i9-7920X CPU. We use the reference implementations of the respective authors and only consider the time for creating the similarity maps. For the latency determination, we use 200 random images from LFW, in total 100 image pairs. To get as close as possible to a realistic application for an end user, we also evaluate the time for a run and process the pairs in pairs and not batch-wise. We report the total time over the 100 image pairs as well as the average time per processed pair. For an end-user application scenario, the user (whether an individual or operator) would expect a timely response, as being verified is not his main goal but e.g. crossing a border or getting access to a restricted area/device.

5. Results

In this section, we present the qualitative and quantitative results, as well as the efficiency of our approach compared to the state-of-the-art. We start with a short qualitative analysis to investigate the visualized explanation maps produced by our approach, xSSAB. We then investigate the performance of our approach and state-of-the-art on the novel Patch-LFW benchmark following the protocol introduced in Section 4.3. Last, we present data on the efficiency of our approach and state-of-the-art.

5.1. Qualitative Analysis

To perform the qualitative analysis, we show two example images in Figure 6. More examples are presented in the supplementary material. xSSAB explanation map is more fine-grained than the explanation maps from xFace [23] and xCos [25]. In the FNM pair, our approach correctly identified the inserted patch at the nose as not similar but detected the insert front head as similar, which is visually understandable. The explanation maps based on the xFace methods are pretty similar and identified the mouth region as similar, and also the inserted nose patch as dissimilar. Both xCos maps also identified the mouth region as similar.

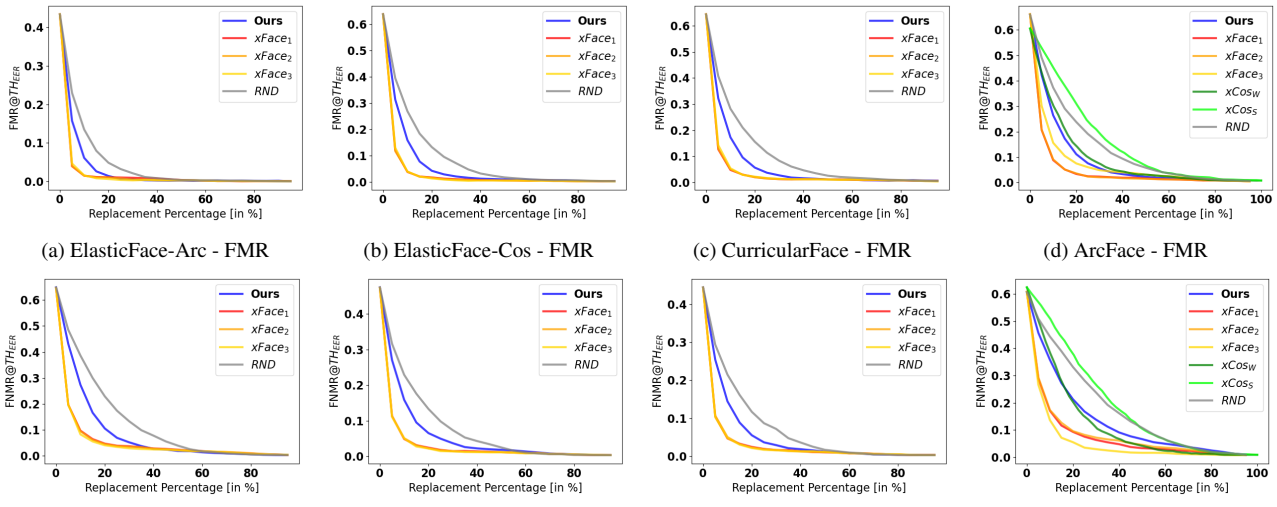
For the FM pair, the explanation maps also look very different. We can observe the same different style as in the FNM pair. All the approaches, besides $xCos_S$ detect the inserted area in the left mouth region as similar. Our approach and also the xFace methods detect the unchanged area between the eyes as being dissimilar. All previous works limited their evaluation to this visual presentation, however, we believe that such solutions should be evaluated in a more statistically significant quantitative manner as we do in the next section.

5.2. Quantitative Analysis - DPR Curve

For a quantitative analysis of the quality of our proposed approach and also to compare it objectively with state-of-the-art, we utilize our proposed Patch-LFW and perform the decision-based patch replacement as described in Section 4.3. The results for the different approaches on the four utilized models for both error rates, FMR and FNMR are shown in Figure 5, and the AUC values are provided in Table 2.

	ElasticFace-Arc		ElasticFace-Cos		CurricularFace		ArcFace	
AUC	FMR	FNMR	FMR	FNMR	FMR	FNMR	FMR	FNMR
$xFace_1$	1.76	5.12	2.96	2.96	3.49	2.84	4.61	7.25
$xFace_2$	1.67	5.06	2.88	2.94	3.51	2.85	4.58	7.83
$xFace_3$	1.63	4.81	2.92	2.87	3.62	2.80	6.86	5.31
$xCos_W$	-	-	-	-	-	-	9.27	11.60
$xCos_S$	-	-	-	-	-	-	15.28	18.74
Ours	2.57	8.04	5.26	5.34	5.85	4.74	8.34	12.64
RND	4.06	11.99	8.203	7.37	9.25	6.73	12.96	17.01

Table 2. The AUC values for the curves are shown in Figure 5. A lower AUC indicates better explainability performance. Our proposed approach beats the random approach and the $xCos_S$ approach on all models in terms of FMR and FNMR. On the ArcFace model, our training-free approach is competitive with the $xCos_W$ approach, which has to be trained in advance. The $xFace$ approaches perform better than our approach but are not efficient, as evaluated and discussed in Section 4.5



(a) ElasticFace-Arc - FMR (b) ElasticFace-Cos - FMR (c) CurricularFace - FMR (d) ArcFace - FMR

(e) ElasticFace-Arc - FNMR (f) ElasticFace-Cos - FNMR (g) CurricularFace - FNMR (h) ArcFace - FNMR

Figure 5. Decision-based Patch Replacement (DPR) curves on Patch-LFW: The error in terms of FMR and FNMR at the fixed threshold th_{EER} determined on LFW, based on the percentage of replaced pixels. Our approach outperforms the random baseline (RND) on all errors and on all models. Our approach shows to be applicable to a wide range of diverse FR models. Even though the $xFace$ methods show slightly better performance, they are not as efficient as discussed in Section 4.5. On the Arcface model, our approach outperforms the pre-trained $xCos_S$ approach and is comparable to pre-trained $xCos_W$, while being training-free. We did not evaluate the $xCos$-based approaches on other models, since the pre-trained module did not provide competitive results on the other models, proofing that it is necessary to train the $xCos$ module before. The AUC values are provided in Table 2.

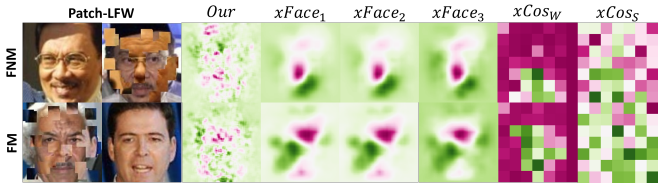


Figure 6. Visualization of the explanation maps of an FNM and an FM pair of Patch-LFW. Our example and the xFace explanation maps are based on the Elasticface-Arc model, the xCos maps are created using the pre-trained module on Arcface. Our approach can detect the inserted patch at the nose as dissimilar to the reference image in the FNM pair. Furthermore, it identified the similar-looking but inserted forehead as similar. In the explanation for the FM, the unchanged area around the eyes is correctly identified as dissimilar.

The results show, that on all models and for both errors, our proposed xSSAB approach outperforms the random baseline (e.g. AUC-FMR of 2.57 to AUC-FMR of 4.06 on random). On the ArcFace model, our training-free efficient approach is competitive to the pre-trained $xCos_W$ approach (FMR-AUC of 8.34 to 9.27 and FNMR-AUC of 12.64 to 11.60) and outperforms the pre-trained $xCos_S$ approach . The three black-box based $xFace$ methods outperform our solution, but they require excessive computing power as we will demonstrate later. For the different versions of $xFace$, no clear performance difference can be observed in our quantitative analysis (e.g. ElasticFace-Arc FMR-AUC of 1.76,1.67 and 1.63). Our efficient approach shows good performance independent from the utilized underlying FR model. Furthermore, it can be observed, that the detection of the inserted patches on the genuine pairs is easier for the applied methods, as the performance regarding FMR, in general, is better than the performance regarding FNMR. This can be also observed in the AUC values, which are generally higher.

To the best of our knowledge, with this evaluation, we provided the first quantitative analysis of explainable face verification methods. Using the newly proposed Patch-LFW with the decision-based patch replacement allows us to evaluate the quality of the similarity and dissimilarity maps in a quantified way, even though Patch-LFW is artificially created. An explainable face verification approach, that is capable to correctly explain non-artificial triggers of verification decisions, should also correctly identify artificially added clues.

5.3. Efficiency

To evaluate the efficiency, we observed the time needed to process and create a single explanation map of a single pair and also to process 100 pairs in total and create the corresponding explanation maps. The times are reported in Table 3. We can observe, that the $xFace_x$ methods are slow,

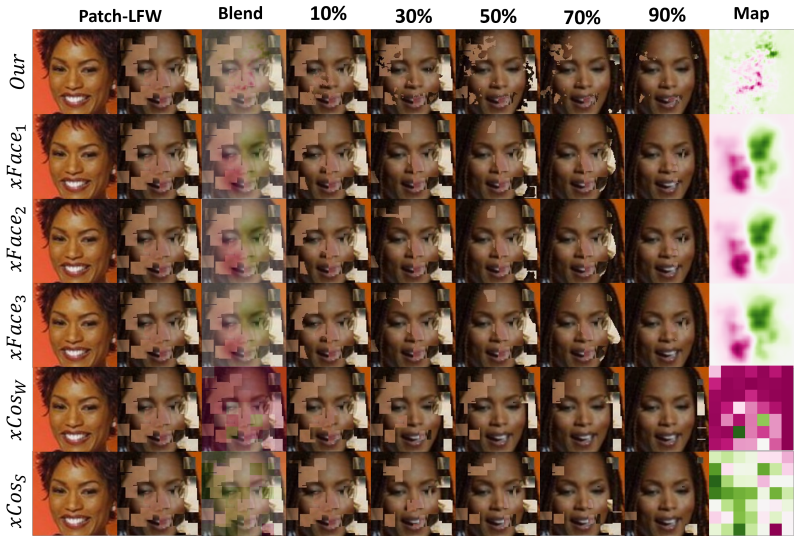


Figure 7. Visualization of the Replacement: Depending on the similarity or dissimilarity, the most influential pixels are replaced with the original pixels from the LFW dataset. Our example and the xFace explanation maps are based on the Elasticface-Arc model, the xCos maps are created using the pre-trained module on Arcface. We can observe, that different areas of the face have been identified as important by the different approaches. Our approach correctly detected the inserted patches at the lower left face side, shortly above the mouth.

Approach	T_{Mean}	T_{Total}
$xFace_1$	12.49s	1249.69s
$xFace_2$	12.53s	1253.31s
$xFace_3$	12.38s	1238.87s
$xCos$	0.11s	11.18s
Ours	0.24s	24.30s

Table 3. Evaluation in terms of latency: The average time needed to create an explanation map, as well as the total time needed to create 100 different ones. Our approach is much faster than the other training-free approaches, $xFace$. $xCos$ is faster but has to be trained in advance, which is not represented here and performs worse.

as they treat the FR model as a black-box and need to manipulate the input multiple times to create meaningful outcomes. In total, 1,130 forward passes are made per image using the proposed parameters by the authors [23], leading to 2,260 forward passes for a single explanation map, which is time-consuming at not very efficient (around 12s per single map). The $xCos$ approach is more efficient in terms of creating explanation maps (0.11s per explanation map) but has to be trained in advance. Our training-free approach is much faster than $xFace_x$ (only 0.24s per explanation map) as it only requires one forward pass and one backward pass per image. It also outperforms $xCos$ as demonstrated earlier in this section.

6. Conclusion

In this work, we proposed xSSAB, an explainable face verification solution based on the backpropagation of similarity score arguments. In our approach, the positive and negative impact of the features on the similarity score based on the FR systems' decision is utilized to obtain visualization of the impact of small, fine-grained face regions on the final matching or non-matching decisions. Our approach efficiently produces fine-grained explanation maps that highlight similar and dissimilar areas as we showed in our experiments and in comparison with two state-of-the-art approaches. To quantitatively evaluate our and other approaches in the field of explainable face verification systems for the first time, we also introduced Patch-LFW. Patch-LFW is a benchmark that is based on LFW [14] and has been artificially manipulated to provide more FMs and FNMs by adding patch-wise clues from another or the same identity. An explainable face verification system can now be evaluated based on the Decision-based Patch Replacement (DPR) curve, which represents how good an explainability solution does its task. In an evaluation of latency, we also showed the efficiency of our approach, which is also training-free in comparison to black-box approaches, which are time-consuming by design. As FR systems become more ubiquitous in our daily lives and the precise workings of highly complex deep-learning-based models remain difficult to comprehend, increasing explainability, interpretability, and transparency will continue to be important, especially in the biometric area, as personal and private data is processed.

Acknowledgement This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science, and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. This work has been partially funded by the German Federal Ministry of Education and Research through the Software Campus Project.

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. 1

[2] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial FC: training 10 million identities on a single machine. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, BC, Canada, October 11-17, 2021*, pages 1445–1449. IEEE, 2021. 1

[3] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognit.*, 120:108102, 2021. 1

[4] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022. 1, 2, 4, 6

[5] Fadi Boutros, Patrick Siebke, Marcel Klemt, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Pocketnet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022. 1

[6] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5709–5718. Computer Vision Foundation / IEEE, 2020. 2

[7] Kai Chen, Taihe Yi, and Qi Lv. Fast and reliable probabilistic face embeddings based on constrained data uncertainty estimation. *Image Vis. Comput.*, 121:104429, 2022. 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 2, 4, 5, 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[10] Biying Fu and Naser Damer. Explainability of the implications of supervised and unsupervised face image quality estimations through activation map variation analyses in face recognition models. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 349–358. IEEE, 2022. 2

[11] Biying Fu and Naser Damer. Towards explaining demographic bias through the eyes of face recognition models. In *IEEE International Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–10. IEEE, 2022. 2

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 1, 4

[14] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 4, 5, 6, 8

[15] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5900–5909. Computer Vision Foundation / IEEE, 2020. 2, 4, 6

[16] Marco Huber, Meiling Fang, Fadi Boutros, and Naser Damer. Are explainability tools gender biased? a case study on face presentation attack detection. In *2023 31st European Signal Processing Conference (EUSIPCO)*, pages 945–949, 2023. 2

[17] Marco Huber, Philipp Terhörst, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper. Stating comparison score uncertainty and verification decision confidence towards transparent face recognition. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 506. BMVA Press, 2022. 1, 2

[18] Marco Huber, Philipp Terhörst, Florian Kirchbuchner, Arjan Kuijper, and Naser Damer. Uncertainty-aware comparison scores for face recognition. In *11th International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*, pages 1–6. IEEE, 2023. 2

[19] ISO/IEC JTC1 SC37 Biometrics. International Organization for Standardization, 2021. 5

[20] Anil K. Jain, Debayan Deb, and Joshua J. Engelsma. Biometrics: Trust, but verify. *IEEE Trans. Biom. Behav. Identity Sci.*, 4(3):303–323, 2022. 1

[21] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584, 2017. 2

[22] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18729–18738. IEEE, 2022. 4

[23] Martin Knoche, Torben Teepe, Stefan Hörmann, and Gerhard Rigoll. Explainable model-agnostic similarity and confidence in face verification. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV 2023 - Workshops, Waikoloa, HI, USA, January 3-7, 2023*, pages 1–8. IEEE, 2023. 1, 2, 4, 5, 6, 8

[24] Jan Niklas Kolf, Fadi Boutros, Jurek Elliesen, Markus Theuerkauf, Naser Damer, Mohamad Alansari, Oussama Abdul Hay, Sara Alansari, Sajid Javed, Naoufel Werghi, Klemen Grm, Vitomir Struc, Fernando Alonso-Fernandez, Kevin Hernandez-Diaz, Josef Bigün, Anjith George, Christophe Ecabert, Hatef Otroshi-Shahreza, Ketan Kotwal, Sébastien Marcel, Iurii Medvedev, Bo Jin, Diogo Nunes, Ahmad Hassanpour, Pankaj Khatiwada, Aafan Ahmad Toor, and Bian Yang. Efar 2023: Efficient face recognition competition. *CoRR*, abs/2308.04168, 2023. 1

[25] Yu-Sheng Lin, Zhe-Yu Liu, Yu-An Chen, Yu-Siang Wang, Ya-Liang Chang, and Winston H. Hsu. Xcos: An explainable cosine metric for face verification task. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(3s), nov 2021. 1, 2, 5, 6

[26] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on LFW with gaussianface. In Blai Bonet and Sven Koenig, editors, *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3811–3819. AAAI Press, 2015. 2

[27] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021. 1, 2

[28] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021. 4

[29] Domingo Mery. True black-box explanation in facial analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1595–1604. IEEE, 2022. 1, 2

[30] Domingo Mery and Bernardita Morris. On black-box explanation for face verification. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1194–1203. IEEE, 2022. 1, 2

[31] Pedro C. Neto, Tiago Gonçalves, João Ribeiro Pinto, Wilson Silva, Ana F. Sequeira, Arun Ross, and Jaime S. Cardoso. Explainable biometrics in the age of deep learning, 2022. 1

[32] Pedro C. Neto, João Ribeiro Pinto, Fadi Boutros, Naser Damer, Ana F. Sequeira, and Jaime S. Cardoso. Beyond masks: On the generalization of masked face recognition models to occluded face recognition. *IEEE Access*, 10:86222–86233, 2022. 4

[33] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[34] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: randomized input sampling for explanation of black-box models. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 151. BMVA Press, 2018. 5

[35] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. 2

[36] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019. 2

[37] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. 3

[38] Philipp Terhörst, Marco Huber, Naser Damer, Florian Kirchbuchner, Kiran B. Raja, and Arjan Kuijper. Pixel-level face image quality assessment for explainable face recognition. *IEEE Trans. Biom. Behav. Identity Sci.*, 5(2):288–297, 2023. 2

[39] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2022. 4

[40] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 111–119. Computer Vision Foundation / IEEE, 2020. 2

[41] Dan Zeng, Raymond N. J. Veldhuis, and Luuk J. Spreeuwers. A survey of face recognition techniques under occlusion. *IET Biom.*, 10(6):581–606, 2021. 4