MorDeephy: Face Morphing Detection via Fused Classification

Iurii Medvedev[1] [a], Farhad Shadmand[1] [b] and Nuno Gonçalves[1,2] [c]

[1]*Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal*

[2]*Portuguese Mint and Official Printing Office (INCM), Lisbon, Portugal*

iurii.medvedev@isr.uc.pt

Keywords: Face Morphing Detection, Face Recognition, Deep Learning, Convolutional Neural Networks, Classification.

Abstract: Face morphing attack detection (MAD) is one of the most challenging tasks in the field of face recognition nowadays. In this work, we introduce a novel deep learning strategy for a single image face morphing detection, which implies the discrimination of morphed face images along with a sophisticated face recognition task in a complex classification scheme. It is directed onto learning the deep facial features, which carry information about the authenticity of these features. Our work also introduces several additional contributions: the public and easy-to-use face morphing detection benchmark and the results of our wild datasets filtering strategy. Our method, which we call MorDeephy, achieved the state of the art performance and demonstrated a prominent ability for generalizing the task of morphing detection to unseen scenarios.

1   INTRODUCTION

The development of deep learning techniques in the last decades allowed to achieve evident advances in the area of face recognition. However, evolved and sophisticated techniques for performing the presentation attacks continue to appear, which require the development of new protection solutions.

Face morphing is one such image manipulating technique. It is usually performed by blending several (usually two) digital face images and allows to match different persons with this synthetic image that contains characteristics from both faces. While being simple to implement, face morphing poses the security risks of issuing an identification document that may be validated for two or more persons. Presentation attacks with face morphing usually can be hardly detected by humans, which usually perform poorly in matching unfamiliar faces on photos of ID and travel documents (Medvedev et al., 2020) and by face recognition software in ABC (automatic border control) systems (Ferrara et al., 2014).

In the last years, face morphing has become a matter of research interest in academia (NIST, 2020) and industry (Research and Service, 2020). Morphing detection methods in facial biometric systems may be distinguished into two pipelines depending on the processing scenario. In *no-reference* morphing attack

[a] [iD] https://orcid.org/0000-0003-2372-9681

[b] [iD] https://orcid.org/0000-0003-4399-4845

[c] [iD] https://orcid.org/0000-0002-1854-049X

detection algorithm receives a single image, where morphing is detected. In practice, these methods are directed to mitigate risks related to the false acceptance of manipulated images in the *enrollment* process. The authentic document, which is generated with a successfully accepted forged image, may further help to deceive the face recognition system.

The *differential* morphing detection implies additional live data acquisition from an authentication system which gives the reference information for the detection algorithm. This scenario usually takes place while passing an Automated Border Control (ABC) system, when the recently enrolled image (which is already accepted and printed on the ID Document) is tested against morphing detection.

First morphing detection solutions relied on the behaviour of local image characteristics (like texture, noise). Recent approaches usually employ deep learning computer vision tools. However, many of these methods utilize a straightforward learning strategy that is limited by binary classification or contrast learning, which in our opinion is not optimal for a task of face morphing detection and may lead to various convergence problems.

In this work, we introduce a novel deep learning method for single image face morphing detection, which incorporates sophisticated face recognition tasks and implies utilizing a combined classification scheme (discussed in Section 3). We propose the new strategy of face morphs classification, which is used for the purpose of face morphing detection.

This novel approach leads to the usage of CNN coupling, which is accompanied by a sophisticated labeling strategy and a specific data mining technique.

Also, we develop the public face morphing detection benchmark, which is designed to be adaptive to the developer needs and at the same time to be simple for comparison of algorithms of different developers. As an additional contribution, we introduce the results of our datasets filtering strategy (image name lists), which is described in Section 4.1.

Regarding the limitations of the work, it is important to note that at the current stage we focus on single image morphing detection. Also, we do not take into account redigitalized face images (by printing/scanning), however we demonstrate that our method allow to generalize the detection onto this case. At the same time, we are limited to utilizing landmark-based methods for performing face morphing. GAN (Generative adversarial Network) based methods require large computational resources (namely for projecting images to latent space) and at the same time, face recognition systems are less vulnerable to presentation attacks with GAN morphs, rather than to landmark-based morphs (Venkatesh et al., 2020). However, we intend to cover those limitations in further research.

2 RELATED WORK

To introduce our methodology, we need to discuss recent advances in face morphing, face morphing detection (focusing on the no-reference scenario) and face recognition.

2.1 Face Morphing

The generic pipeline of creating face morph from original images includes the following steps: face features extraction → features averaging → generating morphed image from averaged features → optional restoring image context (namely background).

Landmark based approaches, first introduced by Ferrara *et al.* (Ferrara et al., 2014), follow this pipeline straightforwardly in the image spatial domain by the face landmark alignment, image warping and blending. Different reported morphing algorithms employ variations of this strategy (Laboratory, 2018; LLC, 2010; Satya, 2016).

With recent advances in generative deep learning approaches, several face morphing methods, which utilize deep latent feature domain, were proposed.

The above face morphing pipeline may adapt various deep learning tools like variational autoencoders (VAE) (Damer et al., 2018) or generative adversarial networks (GANs) (Venkatesh et al., 2020; Zhang et al., 2020).

2.2 Face Morphing Detection

Single image (no-reference) face morphing detection algorithms usually utilise local image information and image statistics.

Various morphing detection approaches employ Binarized Statistical Image Features (BSIF) (Raghavendra et al., 2016), Photo Response Non-Uniformity (PRNU), known as sensor noise (Debiasi et al., 2018; Scherhag et al., 2019), texture features (Ramachandra et al., 2019), local features in frequency and spatial image domain (Neubert et al., 2019) or complex combination of these features (Lorenz et al., 2021; Scherhag et al., 2017).

Several deep learning approaches for no-reference case were proposed. For face morphing detection these approaches usually follow binary classification of pretrained face recognition features (Raghavendra et al., 2017), which may be finetuned (Seibold et al., 2017; Ferrara et al., 2021) or utilized in a combination with local texture characteristics (Wandzik et al., 2018). Damer *et al.* (Damer et al., 2021) introduced a better regularized strategy for morphing detection by replacing the trivial binary classification with pixel-wise supervision. Aghdaie *et al.* (Aghdaie et al., 2021) adopted the attention mechanism which is controlled by wavelet decomposition.

Differential face morphing detection is a less challenging task and security risks in this scenario indeed may be combated by increasing the discriminability of face deep representation, which is utilized for recognition.

Several approaches for differential detection was recently proposed. Scherhag *et al.* (Scherhag et al., 2020) followed the classification of pretrained deep features in differential scenario. Borghi *et al.* (Borghi et al., 2021b) performed differential morphing detection by finetuning the pretrained networks in a complex setup with identity verification and artifacts detection blocks. Rather different approach to the differential scenario was introduced by Ferrara *et al.* (Ferrara et al., 2018) who proposed an approach to revert morphing by retouching the testing face image with a trusted live capture to reveal the identity of the legitimate document owner.

In comparison to the considered approaches, we propose to focus our method on learning the authenticity of deep face features, regularizing the morphing detection with a delicate face recognition task.
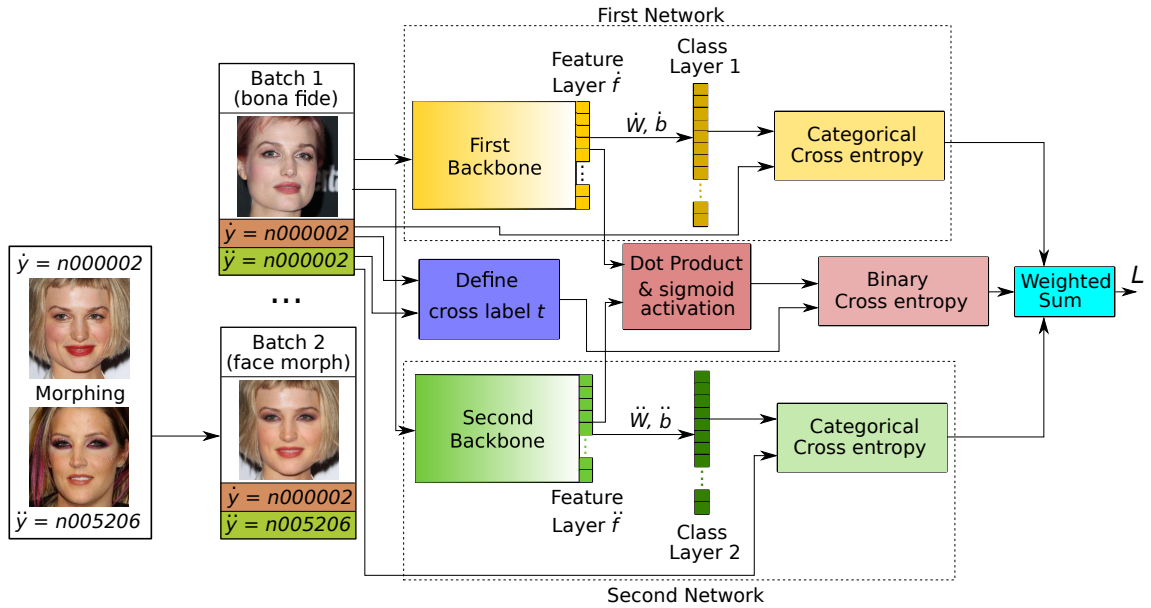
Figure 1: Schematic of the proposed method. For simplicity of visualization batch contains a single image. Labels $\dot{y}$ and $\ddot{y}$ are indicated by names, when the real setup utilizes their numerical index value, which is encoded to one-hot vector.

2.3 Face Recognition

Modern face recognition approaches rely on deep learning tools, which give the ability to learn highly discriminative features themselves from unconstrained images. Among several techniques to perform the tasks of extracting features, the convolutional neural network (CNN) is one of the most efficient for the pattern recognition problems (Russakovsky et al., 2015).

There are several strategies for approaching face recognition via deep learning. However, all of them are focused on extracting low-dimensional face representation (deep face features) and increasing the discriminative power of that representation.

Metric learning methods are directed on optimizing the face representation itself through the contrast of match/non-match pairs (Chopra et al., 2005; Schroff et al., 2015). However, for reliable convergence, these methods require enormously large datasets and sophisticated sample mining techniques.

Another concept (which we indeed follow in our work) is learning face representation implicitly via a closed-set identity classification task. Deep networks in these methods encapsulate face representation in the last hidden layer and usually adopt softmax loss and its modifications for classification (Sun et al., 2014; Sun et al., 2014; Sun et al., 2015).

Improvement of the performance in this approach was achieved by various techniques for increasing intra-class compactness and maximizing inter-class discrepancy. For example, by applying additional regularization for pushing intra-class features to their centre (Wen et al., 2016), or by introducing several kinds of marginal restrictions for inter-class variance (Liu et al., 2017; Wang et al., 2018; Deng et al., 2019; Sun et al., 2020).

Several recent works were directed onto investigating sample specific learning strategies, which are controlled by sample quality (Tremoço et al., 2021), hardness (Zeng et al., 2020; Huang et al., 2020), data augmentation (Shi et al., 2020) or even by treating facial representation in distributional manner (by specifying sample *uncertainty*) (Shi and Jain, 2019).

In our work, we consider face morphing detection from the perspective of face recognition. In the case of following the approach via identity classification, face morphing introduces a problem, since a face morph image indeed belongs to several identities, which leads to the ambiguity of proper class labeling. In this work, we address this issue (Section 4.2) in search of the method for single image morphing detection.

3 METHODOLOGY

In this section, we describe our technique for single image morphing detection via deep learning.

In our research, we intended to design a setup, which will allow to learn high-level deep features, that also carry information about their authenticity. We defined two requirements to our setup. First, each in-

put image must be classified explicitly and unambiguously. Second, the face morphing detection decision is made by the behavior of deep face features.

This resulted in the schematic that includes two backbone CNN based networks that are trained in a similar manner but biased in a way to discriminate morphed and bona fide images. Namely, our idea implies training two parallel networks, which consider bona fide samples similarly and morphed samples differently (see Fig. 1). We point out that these networks do not share the same weights and we do not intend to train them in a contrastive manner (by matching positive and negative pairs).

In our setup both networks learn high-level features via classification tasks, which are different in terms of identity labeling of face morphs. *First Network* labels them by the original identity from the first source image, the *Second Network* - by the second original label.

The extracted features are also explicitly compared by similarity metric (which is the dot product due to the softmax properties) and the result is classified according to the ground truth authenticity label of the image (bona fide/morph). Indeed the variance of the behavior of the networks is used to make the detection decision.

The identity classification parts of the training scheme act as a regularization that retains the facial discriminability of feature layers. That is why for identity classification we utilize a standard softmax, which allows easier convergence in comparison with its modifications (like ArcFace(Deng et al., 2019)).

Following the common formulation of softmax, our training process is regularized by the losses:

$$L_1 = -\frac{1}{N} \sum_i^N \log\left(\frac{e^{\dot{W}_{\dot{y}_i}^T \dot{f}_i + \dot{b}_{\dot{y}_i}}}{\sum_j^C e^{\dot{f}_{\dot{y}_j}}}\right) \qquad (1)$$

$$L_2 = -\frac{1}{N} \sum_i^N \log\left(\frac{e^{\ddot{W}_{\ddot{y}_i}^T \ddot{f}_i + \ddot{b}_{\ddot{y}_i}}}{\sum_j^C e^{\ddot{f}_{\ddot{y}_j}}}\right), \qquad (2)$$

where $\{\dot{f}_i, \ddot{f}_i\}$ denote the deep features of the $i-th$ sample, $\{\dot{y}_i, \ddot{y}_i\}$ are the numerical class indexes of the $i-th$ sample, $\{\dot{W}, \ddot{W}\}$ and $\{\dot{b}, \ddot{b}\}$ are weights and biases of last fully connected layer (respectively for the $\{First, Second\}$ networks). N is the number of samples in a batch and C is the total number of classes.

The target driver of the training process explicitly discriminates morph/non-morph images. The dot product of backbones outputs indicates the morphing detection score. It is activated by sigmoid function, and the corresponding loss is defined as binary cross-entropy:

$$L_3 = -\frac{1}{N} \sum_i^N t \log \frac{1}{1+e^{-D}} + (1-t) \log\left(1 - \frac{1}{1+e^{-D}}\right), \qquad (3)$$

where closs-label t is computed by a comparison of input class labels:

$$t = abs(sgn(\dot{y}_i - \ddot{y}_i)), \qquad (4)$$

and D is a dot product of high level features extracted by *First* and *Second* backbones:

$$D = \dot{f} \cdot \ddot{f} \qquad (5)$$

The result loss for optimization is combined as a weighted sum:

$$L = \alpha_1 L_1 + \alpha_2 L_2 + \beta L_3 \qquad (6)$$

By minimizing this loss in the fused classification setup, we learn the discriminative facial features that are explicitly regularized for morphing detection. The optimal values of weighting coefficients α_1, α_2 and β will be obtained empirically in the Section 6.1.

At the testing stage, the identity classification parts of the network are redundant and may be removed from the setup. The morphing detection decision is made by thresholding the scalar product of the backbones outputs.

4 DATASETS

The proposed methodology requires the large labeled face dataset with an accompaniment of morphed images of identities from this dataset.

The academic community still doesn't have public ID document compliant datasets which are large enough for efficient training of modern deep networks (as an example, one of the largest FRGC_V2(Phillips et al., 2005) contains only ∼50k images and ∼500 identities). That is why our strategy for this work is to utilize the wild dataset which is filtered by the criteria of *suitability for face morphing*. Conceptually this approach indeed is not novel and was recently utilized in face morphing research (Damer et al., 2019; Damer et al., 2021). In this work, we introduce a technique for semi-automatic wild dataset filtering for our method.

As a source wild dataset we use VGGFace2(Cao et al., 2018)(∼3M images, ∼9k classes, ∼360 samples per class, License - CC BY-SA 4.0) due to large average number of samples per identity (in comparison to other popular wild face datasets like CASIA-WebFace(Yi et al., 2014), MS-Celeb-1M(Guo et al., 2016; Jin et al., 2018), Glint360K(An et al., 2021), WebFace260M(Zhu et al., 2021)) in order to have enough samples per identity after filtering.
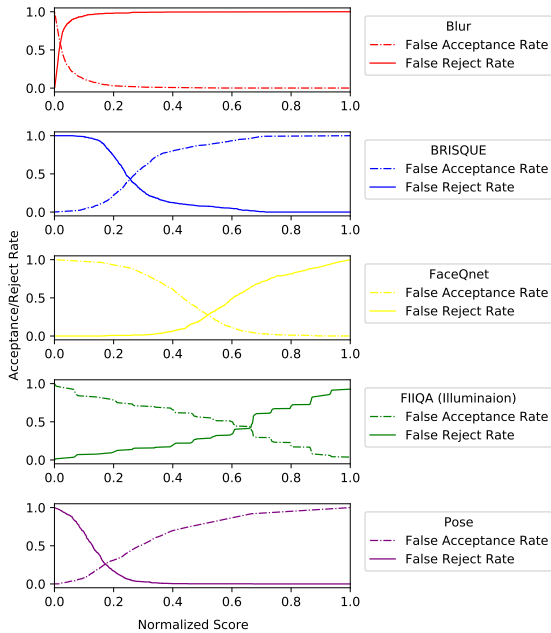
Figure 2: FAR and FRR for result manual quality labeling of random samples from VGGFace2.

4.1 Wild Dataset Filtering

Our dataset filtering strategy is based on a thresholding by quality metrics. Following Tremoço *et al.* (Tremoço et al., 2021) we used a set of quality scores for labeling the face images in the dataset: Blur (Bansal et al., 2016), FaceQNet (Hernandez-Ortega et al., 2019), BRISQUE (Mittal et al., 2012), Face Illumination (Zhang et al., 2017) and Pose (Ruiz et al., 2018). This set of scores allow to discriminate and select samples by their natural quality (Blur, BRISQUE), ID documents suitability (Face-QNet), face image acquisition parameters (Illumination, Pose).

Next, we randomly select samples and manually label them with a binary value (accept/reject). This acceptance is defined by the criteria of suitability for application in face morphing (namely by user's choice). In our setup, we assure that samples are selected distributively across the quality scores values. Namely, we split the total quality score range into a set of sub-ranges and define the minimum number of samples to be selected from each sub-range. By proceeding around 4k images in our setup, we harvest the dependency of FAR (False Acceptance Rate) and FRR (False Rejection Rate) from quality scores values (see Fig. 2).

The dataset filtering is then performed with joint thresholding by those quality metrics. For each score, we select the threshold at a point of EER (Equal Error Rate). To combine a filtered dataset we collect samples, which have a quality score value higher then the corresponding threshold for each quality metric. As a result we get the *VGGFace2-selected* dataset with the same identity list as original source and around 500k images (see Fig. 3).

Figure 3: Example of VGGFace2 filtering result. Accepted images (green box) and Rejected images (red box).

4.2 Morph Dataset Harvesting Strategy

For application in our method, the filtered wild dataset is needed to be accompanied by a large collection of face morphs. We automatically generate these images with our customized landmark-based morphing approach (with blending coefficient 0.5) (see Fig. 4).

A key requirement for effective learning is to provide unambiguity of proper class labeling in our training method. Namely after generating face morph from two arbitrary samples of the original dataset, the resulting image indeed belongs to both source identities. That is why fully random image pairing (for generating morphs) will result in classification confusion.

To avoid that we utilize the following strategy. First, we separate the total list of identities into two disjoint parts, which are attributed to the *First* and the *Second* networks respectively. Next, to generate a face morph, we randomly pair images from identities of these list halves. Each generated image is then labeled according to the attributed sublist for classification by the *First* and the *Second* networks. Let us note that this labeling is made differently for each morphed image and similarly for bona fide images in both networks (see Fig. 1). That is why this technique, which primarily acts as a regularization, also amplifies the morphing detection performance.

The separation of a dataset to two disjoint identity sets is performed to assure that morphed combinations of a particular identity are classified similarly by the networks. For example, if the dataset include four identities $[A, B, C, D]$, we split it to $[A, B]$ (assigned for the *First* network) and $[C, D]$ (assigned for the *Second* network) and generate possible paired morphs:

[AC], [BC], [AD], [BD], which may be unambiguously classified by both networks by their assigned identity lists. For instance, in this case, the combination [AB] is hardly classified. Indeed we assign half of the identities as main to both networks. Without such separation, the overall classification loses its regularisation sense.

Following the above procedure, we generate *VGGFace2-selected-morph* dataset, which contains around 1M morphed images.
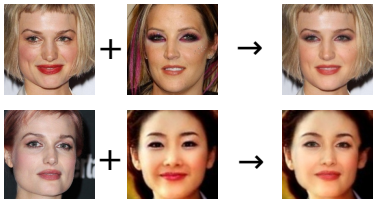

Figure 4: Examples of generated morphs with landmark based approach. Background is restored by one of the source images (chosen randomly).
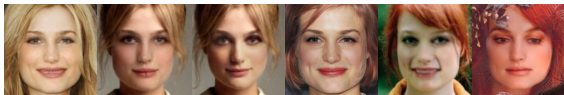

Figure 5: Examples of generated selfmorphs. Images contain blending artefacts but the identity is perceptually retained.

4.3 Selfmorphing

The fully automatic landmark morphing methods may introduce a number of visible artefacts to the generated images (like blending artefacts). That is why without additional regularisation our method will be biased to learning those artefacts, which is not a realistic scenario. Real fraudulent morphs are retouched with the intention to remove any perceptual artefacts.

To address this problem, we utilize *selfmorphs*, which are generated by applying face morphing to images of the same identity. This concept was indeed recently introduced by Borghi *et al.* (Borghi et al., 2021a) and was used for generating images with visible artefacts. Then for removing these artefacts the authors trained the Conditional GAN using original images as a ground truth reference.

In this work, we utilize selfmorph images to focus morphing detection onto the deep face features behaviour, rather than to detecting artefacts. We assume that deep discriminative face features remain after performing selfmorphing. In the proposed method schematic (Fig. 1) we consider selfmorphs as bona fide samples.

We perform a random pairing of samples within each identity from the *VGGFace2-selected* and gen-erate *VGGFace2-selected-selfmorph* dataset, which contains around 500k images (see Fig. 5).

5 BENCHMARKING

There are few public benchmarks for evaluating the performance of morphing detection or morphing resistant algorithms: the NIST FRVT MORPH (NIST, 2022) and FVC-onGoing MAD(Raja et al., 2020; Dorizzi et al., 2009). Both of these benchmarks accept no-reference and differential morphing algorithms, however, they are proprietary and are executed on the maintainer side. Thus they have a number of submission restrictions.

The straightforward metric for evaluating single image morphing detection is the dependency of Bona fide Presentation Classification Error Rate (BPCER) from Attack Presentation Classification Error Rate (APCER) (according to ISO/IEC 30107-3 (International Organization for Standardization, 2017)), which may be plotted as a Detection Error Trade-off (DET) curve.

5.1 Face Morphing Detection Benchmark

For this work, we intend to develop and provide an easy-to-use benchmark, which is to be executed on the developer side (*It will be made public in case of paper acceptance*). The existing public benchmarks provide useful data but usually specify the protocols for the certain software frameworks (Sarkar et al., 2020).

Our benchmark intends to provide the functionality for estimating the morphing detection performance, for generating custom protocols and also for further comparison of the results from different developers with existing protocols. At this stage of our work, we focus on the single image morphing detection with only the usage of public data (however, we assume the possibility of further adapting private datasets).

We generate several protocols for single image morphing detection. Our benchmark is based on the FRGC-Morphs, FRLL-Morphs (Sarkar et al., 2020), AMSL (Neubert et al., 2018) and Dustone datasets(Dustone, 2017). Using this data we combine several benchmark protocols with various types of face morphs:

- *protocol-real* (∼3k morphs(Dustone+AMSL)), which includes morphs with low level of visible blending artifacts, and imitates realistic presentation attacks.
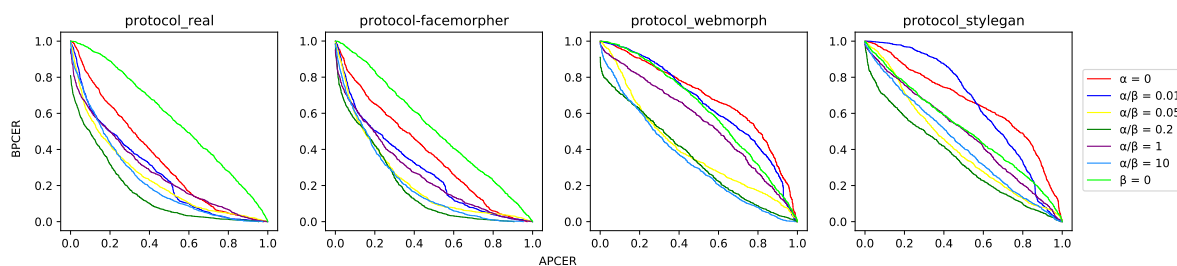
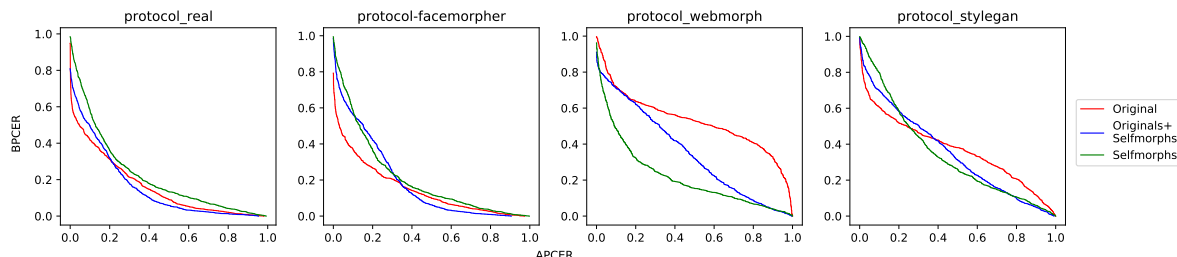Figure 6: Detection Error Trade-off curves for various α/β proportions in different protocols.

Figure 7: Detection Error Trade-off curves for various bona fide images selection in different protocols.

Table 1: APCER@BPCER = (0.1, 0.01) of our method for various α/β proportions and bona fide images selection in different protocols. *Org.* and *SM.* correspond to cases where only original and selfmorph images respectively were chosen as bonafide samples.

Method	$APCER@BPCER = \delta$							
	real		facemorpher		webmorph		stylegan	
	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$
$\alpha = 0$	0.697	0.947	0.756	0.939	0.976	0.995	0.980	0.998
$\alpha/\beta = 0.01$	0.601	0.895	0.651	0.957	0.945	0.992	0.895	0.991
$\alpha/\beta = 0.05$	0.607	0.965	0.502	0.968	0.915	0.999	0.842	0.996
$\alpha/\beta = 0.2$	0.401	0.835	0.431	0.786	0.778	0.979	0.799	0.969
$\alpha/\beta = 1.0$	0.711	0.935	0.670	0.928	0.942	0.995	0.913	0.982
$\alpha/\beta = 10.0$	0.556	0.839	0.513	0.791	0.749	0.923	0.852	0.969
$\beta = 0.0$	0.945	0.996	0.922	0.993	0.954	0.997	0.949	0.997
$\alpha/\beta = 0.2$ Org.	0.481	0.875	0.498	0.876	0.987	0.997	0.915	0.993
$\alpha/\beta = 0.2$ SM.	0.608	0.938	0.568	0.911	0.704	0.986	0.816	0.983

Table 2: BPCER@APCER = (0.1, 0.01) of our method for various α/β proportions and bona fide images selection in different protocols. *Org.* and *SM.* correspond to cases where only original and selfmorph images respectively were chosen as bonafide samples.

Method	$BPCER@APCER = \delta$							
	real		facemorpher		webmorph		stylegan	
	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$
$\alpha = 0$	0.795	0.993	0.781	0.971	0.961	0.998	0.952	0.998
$\alpha/\beta = 0.01$	0.625	0.968	0.638	0.979	0.969	0.997	0.991	1.0
$\alpha/\beta = 0.05$	0.577	0.950	0.598	0.951	0.841	0.989	0.895	0.991
$\alpha/\beta = 0.2$	0.494	0.726	0.562	0.848	0.710	0.822	0.697	0.904
$\alpha/\beta = 1.0$	0.616	0.871	0.628	0.843	0.882	0.963	0.851	0.960
$\alpha/\beta = 10.0$	0.642	0.892	0.604	0.913	0.742	0.931	0.829	0.967
$\beta = 0.0$	0.956	0.998	0.932	0.998	0.971	0.998	0.874	0.986
$\alpha/\beta = 0.2$ Org.	0.417	0.601	0.370	0.595	0.718	0.963	0.605	0.862
$\alpha/\beta = 0.2$ SM.	0.580	0.912	0.587	0.905	0.484	0.842	0.798	0.976

- *protocol-facemorpher* ($\sim$2k morphs), which includes simple morphs with foreground and background artifacts
- *protocol-webmorph* ($\sim$1k morphs), which includes images with background artifacts but the low level of artifacts inside the face contour
- *protocol-stylegan* ($\sim$2k morphs), which includes StyleGan morphs

As bona fide images all our protocols use frontal faces from the following public datasets: FRLL Set(DeBruine and Jones, 2017), FEI (Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil, 2006), AR(Martinez and Benavente, 1998), Aberdeen and Utrecht (School of Natural Sciences University of Stirling, 1998) ($\sim$1.5k images in total).

Since our benchmark utilities are not public yet other developers cannot contribute their results to perform SOTA comparison for this work. That is why we use our protocols only for performing the ablation study and finding the best settings for our approach.

6 EXPERIMENTS AND RESULTS

To analyze the performance of our approach we make several experiments with our method. As backbone networks we use ResNet-50 (He et al., 2016), which are initialized with weights, pretrained on the ImageNet dataset. Followed by pooling and dropout layers each backbone returns 512 deep features. To complete the networks for classification we add the classification layer with the size $= C$ equal to the number of classes in the training dataset. Input images (RGB 3-channel) are aligned and resized to 224$\times$224. In all our experiments we perform training with SGD (Stochastic Gradient Descent) optimizer for 2 epochs, decreasing learning rate from 0.001 to 0.0001. We report the performance by $APCER@BPCER = (0.1, 0.01)$ and $BPCER@APCER = (0.1, 0.01)$.

Our default training dataset is a joined and shuffled concatenation of the above mentioned datasets: *VGGFace2-selected*, *VGGFace2-selected-selfmorph*, and *VGGFace2-selected-morph*.

It is important to note that in all experiments we assured the equal balance between the numbers of morphed and non-morphed (which are bona fide + selfmorphed) images in the training dataset.

6.1 Fused Classification Balance

For effective convergence and further morphing detection, our method requires choosing the proper balance between the elements of the loss function. Namely the balance between α ($= \alpha_1 = \alpha_2$) and β (disbalance of α_1 and α_2 didn't demonstrate any interesting behaviour in our tests). We perform training of our method with different proportional settings also including the ablation of particular parts from the overall loss. Our experiments demonstrate (see Fig. 6 and Tab. 1, 2) that by varying α/β proportion it is possible to achieve some optimal performance of morphing detection in different protocols. Our strategy allows generalizing the detection of morphing even to the images, which are generated with GANs even accounting that this type of morphing is totally unseen in the training.

On the edge case with excluded main loss function driver (namely binary morph/bona fide classification), our method demonstrates the almost random detection decision. At the same time, ablation of the regularization ($\beta = 0$) also leads to bad performance, which we relate with the overfitting on the trivial binary classification learning task.

Summing up, we can conclude that our strategy allows learning such face features which are discriminative by the criteria of authenticity.

6.2 Data Combination Experiments

Further experiments are performed with the selected proportion $\alpha/\beta = 0.2$ in order to understand the impact of selfmorphing for our method.

In comparison to the dataset selection in Section 6.1, where the collection of bona fide samples is split evenly to original and selfmorphs, we test two more options where these particular parts are ablated from the total dataset.

Our results (see Fig. 7 and Tab. 1,2) proves the significant importance of selfmorphs in our strategy. Utilizing selfmorphs at the training stage allows to reduce the emphasis of the detection of facial blending artefacts and shift it to the behavior of the deep feature for generalizing to unseen types of attacks.

6.3 NIST FRVT Morph Results

We have performed the comparison of the results of our method and the state of the art face morphing detection approaches with NIST FRVT MORPH Benchmark (NIST, 2022).

We select the model with $\alpha/\beta = 0.2$ from Section 6.1 as our best model and present results of comparison in several protocols:

- *P1* - Visa-Border (25727 Morphs)
- *P2* - Manual (323 Morphs)
- *P3* - MIPGAN-II (2464 Morphs)

Table 3: Comparison with the state of the art single image morphing detection methods by APCER@BPCER = (0.1, 0.01).

Method	$APCER@BPCER = \delta$									
	P1		P2		P3		P4		P5	
	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$	$\delta = 0.1$	$\delta = 0.01$
(Aghdaie et al., 2021)	**0.172**	0.782	0.965	0.998	0.923	0.991	**0.015**	**0.200**	**0.271**	**0.721**
(Debiasi et al., 2018)	0.994	0.996	**0.049**	0.823	0.994	1.000	1.000	1.000	0.985	0.994
(Ramachandra et al., 2019)	0.659	0.996	0.375	0.990	0.938	0.985	0.159	0.998	0.936	0.996
(Scherhag et al., 2018)	0.986	1.000	1.000	1.000	0.997	1.000	0.996	1.000	0.993	1.000
(Lorenz et al., 2021)	0.844	1.000	0.380	1.000	0.966	1.000	0.819	1.000	0.971	0.995
(Ferrara et al., 2021)	0.982	0.996	0.477	0.999	0.978	1.000	0.037	0.810	0.420	0.777
Ours	0.419	**0.658**	0.434	**0.686**	**0.842**	**0.954**	0.323	0.639	0.499	0.805

- *P4* - Print + Scanned (3604 Morphs)

As bona fide samples all protocols utilize a large collection of 1047389 Bona Fide images. The comparison is performed by the metrics $APCER@BPCER = (0.1, 0.01)$.

6.4 Single Image MAD

We perform the comparison in the target single image morphing detection scenario (see Table 3).

MorDeephy outperforms other techniques in detecting landmark-based morphs and challenging manual morphs and achieves comparable results in other protocols.

Also, our method does not demonstrate bias to a particular morphing generative strategy and has the most stable performance across all protocols in comparison to other approaches.

It is important to note that these results are achieved by utilizing a rather straightforward and simple morphing technique during training (without any adaptation to realistic scenario or modifications for removing artefacts), which proves that our method allows generalizing morphing detection to various unseen generative approaches by focusing on deep face features behavior.

7 CONCLUSION

We introduce a novel deep learning strategy for single image face morphing detection, which implies utilizing a complex classification task. It is directed onto learning the deep facial features, which carry information about the authenticity of these features. Our method achieved the state of the art performance and demonstrated a prominent ability for generalizing the task of morphing detection to unseen scenarios (like GAN morphs and print/scan morphs).

Our work also introduces several additional contributions, which are the public and easy-to-use face morphing detection benchmark and the results of our wild datasets filtering strategy.

In our further work, we will focus on improving the performance by applying more sophisticated morphing techniques during training and on explicit adapting our method to the differential scenario, which will require sophisticated sampling strategies.

REFERENCES

Aghdaie, P., Chaudhary, B., Soleymani, S., Dawson, J. M., and Nasrabadi, N. M. (2021). Attention Aware Wavelet-based Detection of Morphed Face Images. *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8.

An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., and Fu, Y. (2021). Partial FC: Training 10 Million Identities on a Single Machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1445–1449.

Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil (2006). FEI face database. https://fei.edu.br/ cet/facedatabase.html. (accessed: September 1, 2022).

Bansal, R., Raj, G., and Choudhury, T. (2016). Blur image detection using laplacian operator and open-cv. In *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, pages 63–67.

Borghi, G., Franco, A., Graffieti, G., and Maltoni, D. (2021a). Automated artifact retouching in morphed images with attention maps. *IEEE Access*, 9:136561–136579.

Borghi, G., Pancisi, E., Ferrara, M., and Maltoni, D. (2021b). A double siamese framework for differential morphing attack detection. *Sensors*, 21:3466.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1.

Damer, N., Saladié, A. M., Braun, A., and Kuijper, A. (2018). MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–10.

Damer, N., Saladié, A. M., Zienert, S., Wainakh, Y., Terhörst, P., Kirchbuchner, F., and Kuijper, A. (2019). To Detect or not to Detect: The Right Faces to Morph. In *2019 International Conference on Biometrics (ICB)*, pages 1–8.

Damer, N., Spiller, N., Fang, M., Boutros, F., Kirchbuchner, F., and Kuijper, A. (2021). PW-MAD: Pixel-wise Supervision for Generalized Face Morphing Attack Detection. *ArXiv*, abs/2108.10291.

Debiasi, L., Scherhag, U., Rathgeb, C., Uhl, A., and Busch, C. (2018). PRNU-based detection of morphed face images. *2018 International Workshop on Biometrics and Forensics (IWBF)*, pages 1–7.

DeBruine, L. and Jones, B. (2017). Face research lab london set. https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666/3. (accessed: September 1, 2022).

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694.

Dorizzi, B., Cappelli, R., Ferrara, M., Maio, D., Maltoni, D., Houmani, N., Garcia-Salicetti, S., and Mayoue, A. (2009). Fingerprint and On-Line Signature Verification Competitions at ICB 2009. In *Advances in Biometrics : Third International Conference, ICB 2009, Alghero, Italy*, volume 5558.

Dustone, T. (2017). New face morphing database for vulnerability research. https://www.linkedin.com/pulse/new-face-morphing-dataset-vulnerability-research-ted-dunstone. (accessed: September 1, 2022).

Ferrara, M., Franco, A., and Maltoni, D. (2014). The magic passport. *IJCB 2014 - 2014 IEEE/IAPR International Joint Conference on Biometrics*.

Ferrara, M., Franco, A., and Maltoni, D. (2018). Face demorphing. *IEEE Transactions on Information Forensics and Security*, 13(4):1008–1017.

Ferrara, M., Franco, A., and Maltoni, D. (2021). Face morphing detection in the presence of printing/scanning and heterogeneous image sources. *IET Biometrics*, 10.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Proceedings of ECCV*, volume 9907, pages 87–102.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA. IEEE.

Hernandez-Ortega, J., Galbally, J., Fierrez, J., Haraksim, R., and Beslay, L. (2019). FaceQnet: Quality Assessment for Face Recognition based on Deep Learning. In *2019 International Conference on Biometrics (ICB)*, pages 1–8.

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., and Huang, F. (2020). CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

International Organization for Standardization (09 2017). ISO/IEC 30107–3:2017. Information Technology—Biometric Presentation Attack Detection — Part 3: Testing and Reporting. ISO/IEC JTC 1/SC 37 Biometrics.

Jin, C., Jin, R., Chen, K., and Dou, Y. (2018). A community detection approach to cleaning extremely large face database. *Computational intelligence and neuroscience*, 2018.

Laboratory, B. S. (2018). UBO-Morpher. http://biolab.csr.unibo.it/Research.asp. (accessed: September 1, 2022).

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. (2017). SphereFace: Deep Hypersphere Embedding for Face Recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746.

LLC, M. M. (2010). FaceFusion application. www.wearemoment.com/FaceFusion/. (accessed: September 1, 2022).

Lorenz, S., Scherhag, U., Rathgeb, C., and Busch., C. (2021). Morphing attack detection: A fusion approach. In *IEEE Fusion*.

Martinez, A. and Benavente, R. (1998). The AR face database. *Tech. Rep. 24 CVC Technical Report*.

Medvedev, I., Gonçalves, N., and Cruz, L. (2020). Biometric System for Mobile Validation of ID And Travel Documents. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5.

Mittal, A., Moorthy, A. K., and Bovik, A. C. (2012). No-reference image quality assessment in the spatial

domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708.

Neubert, T., Kraetzer, C., and Dittmann, J. (2019). A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD. In *Proceedings of the ACM Workshop*, pages 95–100.

Neubert, T., Makrushin, A., Hildebrandt, M., Kraetzer, C., and Dittmann, J. (2018). Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7(4):325–332.

NIST (2020). International Face Performance Conference. https://www.nist.gov/news-events/events/2020/10/international-face-performance-conference-ifpc-2020. (accessed: September 1, 2022).

NIST (2022). NIST FRVT MORPH. https://pages.nist.gov/frvt/html/frvt_morph.html. (accessed: September 1, 2022).

Phillips, P. J., Flynn, P., Scruggs, T., Bowyer, K., Chang, J. K., Hoffman, K., Marques, J., Min, J., and Worek, W. (2005). Overview of the Face Recognition Grand Challenge. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 947– 954.

Raghavendra, R., Raja, K. B., and Busch, C. (2016). Detecting morphed face images. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7.

Raghavendra, R., Raja, K. B., Venkatesh, S., and Busch, C. (2017). Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1822–1830.

Raja, K., Ferrara, M., Franco, A., Spreeuwers, L., Batskos, I., Wit, F., Gomez-Barrero, M., Scherhag, U., Fischer, D., Venkatesh, S., Singh, J. M., Li, G., Bergeron, L., Isadskiy, S., Raghavendra, R., Rathgeb, C., Frings, D., Seidel, U., Knopjes, F., and Busch, C. (2020). Morphing Attack Detection - Database, Evaluation Platform and Benchmarking. *IEEE Transactions on Information Forensics and Security*, PP:1–1.

Ramachandra, R., Venkatesh, S., Raja, K., and Busch, C. (2019). Towards making morphing attack detection robust using hybrid scale-space colour texture features. In *2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA)*, pages 1–8.

Research, T. C. and Service, D. I. (2020). Image Manipulation Attack Resolving Solutions. https://cordis.europa.eu/project/id/883356. (accessed: September 1, 2022).

Ruiz, N., Chong, E., and Rehg, J. (2018). Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge.

International Journal of Computer Vision (IJCV), 115(3):211–252.

Sarkar, E., Korshunov, P., Colbois, L., and Marcel, S. (2020). Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks. *arXiv preprint*.

Satya, M. (2016). Face Morph Using OpenCV. www.learnopencv.com/face-morph-using-opencv-cpp-python/. (accessed: September 1, 2022).

Scherhag, U., Debiasi, L., Rathgeb, C., Busch, C., and Uhl, A. (2019). Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(4):302–317.

Scherhag, U., Raghavendra, R., Raja, K. B., Gomez-Barrero, M., Rathgeb, C., and Busch, C. (2017). On the vulnerability of face recognition systems towards morphed face attacks. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6.

Scherhag, U., Rathgeb, C., and Busch, C. (2018). Morph Deterction from Single Face Image: a Multi-Algorithm Fusion Approach. In *Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications*, pages 6–12.

Scherhag, U., Rathgeb, C., Merkle, J., and Busch, C. (2020). Deep Face Representations for Differential Morphing Attack Detection. *IEEE Transactions on Information Forensics and Security*, 15:3625–3639.

School of Natural Sciences University of Stirling (1998). Psychological Image Collection of Stirling. http://pics.stir.ac.uk. (accessed: September 1, 2022).

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference CVPR*, pages 815–823.

Seibold, C., Samek, W., Hilsmann, A., and Eisert, P. (2017). Detection of Face Morphing Attacks by Deep Learning. In *Proceedings of Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017*, pages 107–120.

Shi, Y. and Jain, A. (2019). Probabilistic Face Embeddings. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6901–6910.

Shi, Y., Yu, X., Sohn, K., Chandraker, M., and Jain, A. (2020). Towards Universal Representation Learning for Deep Face Recognition. In *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6816–6825.

Sun, J., Yang, W., Xue, J., and Liao, Q. (2020). An Equalized Margin Loss for Face Recognition. *IEEE Transactions on Multimedia*, pages 1–1.

Sun, Y., Chen, Y., Wang, X., and Tang, X. (2014). Deep Learning Face Representation by Joint Identification-Verification. In *NIPS*.

Sun, Y., Wang, X., and Tang, X. (2014). Deep Learning Face Representation from Predicting 10,000 Classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898.

Sun, Y., Wang, X., and Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust.

2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2892–2900.

Tremoço, J., Medvedev, I., and Gonçalves, N. (2021). Qual-Face: Adapting Deep Learning Face Recognition for ID and Travel Documents with Quality Assessment. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–6.

Venkatesh, S. K., Zhang, H., Ramachandra, R., Raja, K. B., Damer, N., and Busch, C. (2020). Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection. *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6.

Wandzik, L., Kaeding, G., and Garcia, R. V. (2018). Morphing Detection Using a General- Purpose Face Recognition System. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1012–1016.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). CosFace: Large Margin Cosine Loss for Deep Face Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham. Springer International Publishing.

Yi, D., Lei, Z., Liao, S., and Li, S. (2014). Learning face representation from scratch. *ArXiv*, abs/1411.7923.

Zeng, D., Shi, H., Du, H., Wang, J., Lei, Z., and Mei, T. (2020). NPCFace: A Negative-Positive Cooperation Supervision for Training Large-scale Face Recognition. *CoRR*, abs/2007.10172.

Zhang, H., Venkatesh, S., Ramachandra, R., Raja, K., Damer, N., and Busch, C. (2020). MIPGAN - Generating Robust and High Quality Morph Attacks Using Identity Prior Driven GAN. *ArXiv*, abs/2009.01729.

Zhang, L., Zhang, L., and Li, L. (2017). Illumination Quality Assessment for Face Images: A Benchmark and a Convolutional Neural Networks Based Model. *Lecture Notes in Computer Science*, 10636 LNCS:583–593.

Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., and Zhou, J. (2021). WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10492–10502.