Unveiling the Two-Faced Truth: Disentangling Morphed Identities for Face Morphing Detection

Eduarda Caldeira*, Pedro C. Neto*, Tiago Gonçalves*, Naser Damer, Ana F. Sequeira, Jaime S. Cardoso

***Abstract*—Morphing attacks keep threatening biometric systems, especially face recognition systems. Over time they have become simpler to perform and more realistic, as such, the usage of deep learning systems to detect these attacks has grown. At the same time, there is a constant concern regarding the lack of interpretability of deep learning models. Balancing performance and interpretability has been a difficult task for scientists. However, by leveraging domain information and proving some constraints, we have been able to develop IDistill, an interpretable method with state-of-the-art performance that provides information on both the identity separation on morph samples and their contribution to the final prediction. The domain information is learnt by an autoencoder and distilled to a classifier system in order to teach it to separate identity information. When compared to other methods in the literature it outperforms them in three out of five databases and is competitive in the remaining.**

***Index Terms*—auto-encoder, biometrics, explainability, face recognition, knowledge, distillation, morphing attack detection, synthetic data.**

I. Introduction

FACE recognition (FR) systems have had large-scale adoption in the most diverse scenarios [1]–[3]. Deep learning (DL) techniques have taken this and other biometric recognition systems towards above-human performance. While it also benefited biometric systems adoption, DL methods led to two problems. First, the approaches that improve the recognition power of these systems are the same to be used to design novel and dangerous attacks [4]. Some attacks can take the form of adversarial noise addition or be developed with FR systems in mind. The latter comprises both morphing [5] and presentation attacks [6]. Besides the attacks, deep learning methods are notorious for their black-box behaviour, which compromises the understanding of both the inner workings of the model and the reasoning behind a decision. Furthermore, FR and face attack detection systems are consistently designed using problem-agnostic tools, which do not leverage domain knowledge. For a wider adoption and to be able to deploy these systems on critical scenarios, it is necessary to guarantee that their reasoning process is, at least to some extent, explained. One can explain a decision using a post-hoc approach [7], or directly interfering with the training behaviour of the model as stated by Neto *et al.* [8].

* These authors contributed equally.

Eduarda Caldeira, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira and Jaime Cardoso are with INESC TEC and University of Porto.

Naser Damer is with the Fraunhofer Institute for Computer Graphics Research IGD and the TU Darmstadt.

Face morphing attacks, which merge two images from different identities into a single image capable of misdirecting the recognition system, have progressed significantly. In other words, this attack aims to increase the number of false positives of the FR system, granting access to two distinct users. When left undetected, the fusion of two images might allow two different people to pass border control with the same passport, for example. Due to this threat, researchers focused their attention on the development of robust morphing attack detection (MAD) systems [4], [9], [10]. Usually, they are designed to detect if the input image is an attack or a *bonafide* sample and do not include any information regarding the fused identities in their training. OrthoMAD [11] aims to learn this information in an unsupervised manner by separating identity information into two orthogonal latent vectors. However, it lacks guarantees regarding the relation between the disentangled information and identity information. The recent blossoming of synthetic data generation methods, such as generative adversarial networks (GAN) and diffusion models, led to the creation of synthetic datasets with a diverse number of identities represented [12], [13]. Although these identities are usually represented only once, it suffices to increase identity diversity.

The work presented in this document builds on top of OrthoMAD premises that it is possible to disentangle information regarding different identities. The first addition is an auto-encoder model trained on the *bonafide* samples to minimize the reconstruction error. The latent vector produced by the encoder is considered to be the prior of the identity information that should be present in the disentangled vectors. We further relax the orthogonality constraint to ensure that the angle between the two identity vectors, in the case of an attack, approximates the angle of the priors of their identities. To achieve this, we leverage the latent vectors of the auto-encoder for both images (before being morphed) and a knowledge distillation strategy. Finally, to further approximate the latent space of both identity vectors we replace the concatenation and classification process with a shared linear layer to be used on both vectors separately. The two predicted scores are fused afterwards.

The main contributions of this work are the following: 1) an unexplored knowledge distillation approach based on the angle of two vectors that represent identity priors; 2) the improvement on the usage of the diverse identity set to regularize the latent spaces and the identity disentanglement; 3) a novel method designed specifically for this domain and with increased transparency regarding its inner workings; 4) an empirical validation and comparison with similar (state-of-the-art) approaches.

ISBN: 978-9-4645-9360-0

This document is divided into five main sections. Besides this introductory section, the following sections include a description of the methodology, an introduction to the databases used for training and evaluation, the experimental setup designed for the experiments, the discussion of the results, and finally the conclusion. The code related to this paper is publicly available in a GitHub repository[1].

II. Methodology

Morphing attacks occur when two distinct identities are fused together, resulting in an image that can trick a face recognition system by containing enough information about both identities. To analyse whether information from two distinct identities is present in an image, we designed a regularisation term based in knowledge distillation (KD). As such, we call IDistill to our proposed method. The overall scheme and architecture of our proposed model is represented on Figure 1.

We start by training an autoencoder to reconstruct *bonafide* images. This autoencoder is responsible for creating a minimalist representation of a face I. Alternatively to the usage of the autoencoder, we could have leveraged a pretrained face recognition system. The decision to follow with the autoencoder yields three reasons: 1) Fang *et al.* [14] has shown a difference in the reconstruction performance from abnormal and normal face images; 2) Besides being large (512-d), the latent vector face recognition systems might not contain all the information necessary for the reconstruction. 3) Encoder-Decoder have been explored for face de-morphing [15]. The proposed autoencoder is based on the U-Net architecture [16] and the size of the latent representation of the image was chosen as 128. As in other reconstruction tasks the autoencoder receives the image I, creates a latent representation u of that image using the encoder, and reconstruct it as $\tilde{I}$ using the decoder network. This approach uses a mean squared error (MSE) loss function (see Eq. 1).

$$L_{auto} = \sum_{i,j} (I_{ij} - \tilde{I}_{ij})^2 \tag{1}$$

The architecture of the morphing classifier is based on a ResNet-18 [17] where the last fully-connected layer is replaced with two fully-connected layers that output two vectors v of size 128 each. Afterwards, a fully-connected layer is used to infer if the vectors contain information of an identity or not, with each vector producing a score (id). The same layer is used for both vectors individually.

$$id = \frac{1}{1 + e^{-W^T v}} : \tag{2}$$

Considering that the produced score holds information regarding the presence of encoded identity information on a vector v, the final prediction for an image I is designed as follows. Given I, the backbone architecture produces v_1 and v_2, which will result in the identity probabilities id_1 and id_2, respectively. The probability of I containing information of two distinct identities is given by $id_1 * id_2$. Consequently, the *bonafide* presentation score, $\tilde{y}$ is given by:

$$\tilde{y} = 1 - id_1 * id_2 \tag{3}$$

For the classification task, we have introduced the Binary Cross-Entropy, L_{BCE} (Equation 4) at the level of the final fused prediction $\tilde{y}$.

$$L_{BCE} = -(y\, log(\tilde{y}) + (1 - y)\, log(1 - \tilde{y})) \tag{4}$$

To ensure that the latent vectors v_1 and v_2 extract identity information and are aligned with the information learnt by the autoencoder, we introduce a knowledge distillation term. For attacks, this term aims to extract vectors from the morphed image that have an angle between them equal to the angle produced by the latent vectors u_1 and u_2 extracted with the encoder from the two images that originated the morphed image (Eq. 7). We are then promoting a proximity between the autoencoder latent space and the morphing classifier latent space while handling attacks. Furthermore, we only consider the angle formed by these vectors, since their identity intensity might be diminished in the morphing process. For *bonafide*, we expect one vector to hold identity information, while the other does not. As such, we designed a term that first selects the vector v with the highest cosine similarity (S_{cos}) to u. With this choice, the proposed term is able to maximize the similarity between u and the selected vector v, while approximating the id of this vector to 1, and the other id to 0 (Eq. 6).

$$Ver_{term} = S_{cos}(v_1, u) > S_{cos}(v_2, u) \tag{5}$$

$$L_{KD_1} = \begin{cases} (1 - id_1)^2 + (id_2)^2 - S_{cos}(v_1, u) & \text{if } Ver_{term} \\ (1 - id_2)^2 + (id_1)^2 - S_{cos}(v_2, u) & \text{otherwise} \end{cases} \tag{6}$$

$$L_{KD_2} = [S_{cos}(u_1, u_2) - S_{cos}(v_1, v_2)]^2 \tag{7}$$

$$L_{KD} = yL_{KD_1} + (1 - y)L_{KD_2} \tag{8}$$

Both losses are incorporated in a single equation as follows:

$$Loss = L_{BCE} + L_{KD} \tag{9}$$

III. Databases

This work builds on top of a proposal by Neto et al. [11], hence, we use the same datasets to train and test our methodology:

1) FRLL: The Face Research London Lab dataset [18] was used to produce the FRLL-Morphs dataset [19], which is frequently used to test morphing attack detection methods. Five different morphing techniques are used in the dataset, including Style-GAN2 [20], [21], WebMorph [22], AMSL [23], FaceMorpher [24], and OpenCV [25]. Each of the five methods uses 204 genuine samples and more than one thousand morphed

[1]https://github.com/NetoPedro/IDistill

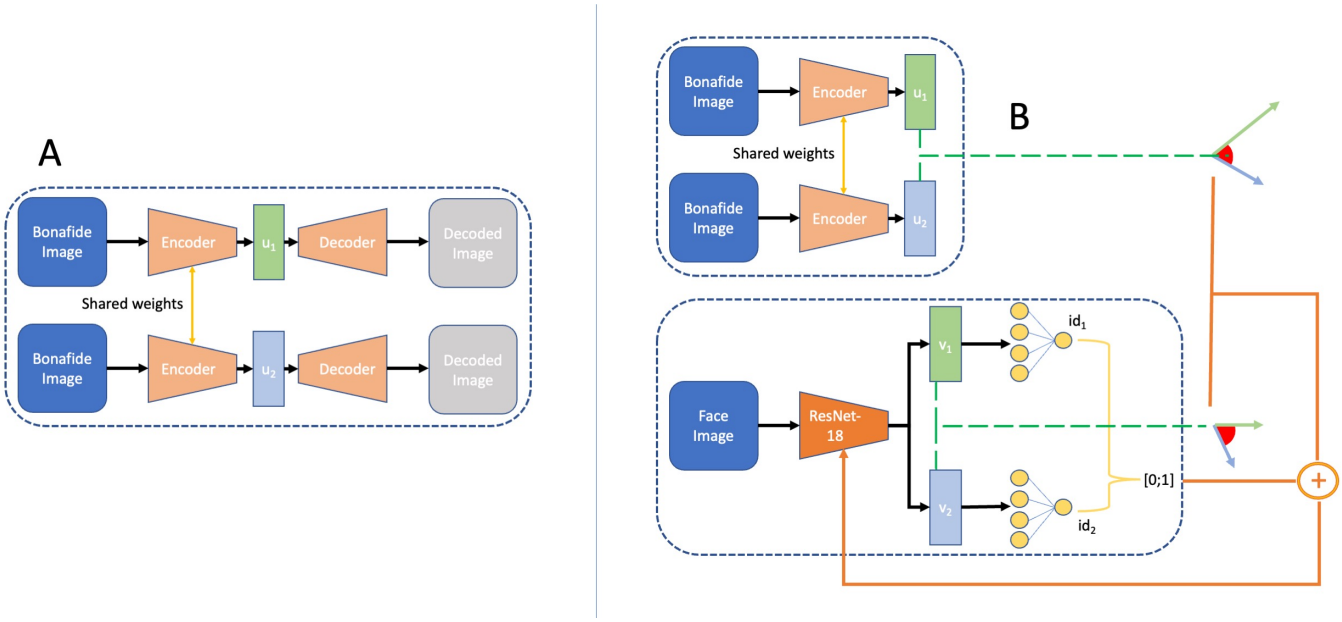

Fig. 1. Overall scheme of the architecture. Part A represents the training of the autoencoder, whereas part B represents the training of the classification system. The orange line represents the backpropagation, the green one represents the calculation of the angle between vectors and the yellow represents the fusion of the two scores. Best viewed in color.

faces made from high-resolution frontal faces. We used this database only for evaluation purposes because it lacks distinct train, validation or test sets.

2) SMDD: The Synthetic Morphing Attack Detection Development (SMDD) [12], is a novel dataset that uses synthetic images to create a dataset of morph and *bonafide* samples. It initially generated 500k images of faces using a random Gaussian noise vector sampled from a normal distribution using the official open-source version of StyleGan2-ADA [20]. Leveraging the quality estimation method known as CR-FIQA [26], 50k of these photos were chosen for analysis because of their high quality, and 25k of them were determined to be the *bonafide* samples. The attack photos were paired with five other attack images at random, and 5k of them were chosen as key morphing images. Next, using the OpenCV [25] method, they were morphed, yielding 15k attack samples. The original 25k images that were used to generate the morphs are also publicly available. This dataset was divided in test and validation sets, on a proportion of 85-15%.

IV. Experimental Setup

The autoencoder was trained for 300 epochs, with a learning rate of 1×10^{-4}, a batch size of 32, and Adam [27] was used as the optimisation algorithm to minimize the MSE loss. It trained exclusively on *bonafide* samples.

The classifier was trained with the joint loss (Eq. 9) utilizing a learning rate of 1×10^{-4}, a batch size of 16 and was optimised with Adam. Furthermore, to align with the autoencoder, both v_1 and v_2 are 128-d. The training utilized the synthetic dataset SMDD, which allowed for this regularization term to utilise the original samples that originated the morphing samples.

To evaluate the performance of the morphing detection, we evaluated our algorithm using different metrics, commonly used in the literature: the Attack Presentation Classification Error Rate (APCER) (i.e., morphing attacks classified as *bonafide*); and the *Bonafide* Presentation Classification Error Rate (BPCER) (i.e., the *bonafide* samples that are classified as morphing attacks). We evaluated these metrics at two different fixed APCER values (1.0% and 20.0%). The equal error rate (EER), which is the BPCER and APCER at the decision thresholds where they are the same, was also evaluated.

V. Results and Discussion

The literature on face morphing attack detection is large, however, is also disperse. In other words, the datasets used for benchmarking and training are not always the same, and as such, direct comparisons are not trivial. The combination of FRLL and SMDD has been found in at least two different documents in the literature. The first [12] introduces the SMDD datasets and evaluate three different methods from the literature: Inception [28], PW-MAD [9] and MixFacenet [29]. Their results vary and there is not one that beats the others cosistently across the different FRLL morphing methods. Afterwards, OrthoMAD was also evaluated using the exact same protocol [11] and achived state-of-the-art results on three out of the five morphing approaches.

Since we follow the protocol introduced by Damer *et al.* [12], the comparison between our method and the ones in the literature focuses on the above mentioned approaches. The results of our method, IDistill, are displayed in Table I. As seen, IDistill has been able to surpass MixFacenet and Inception in all the test databases, and PW-MAD in four out of five databases. OrthoMAD has better results on two databases. A careful analysis of the results highlights an important notion

TABLE I
RESULTS COMPARISON WITH FOUR MODELS PUBLISHED IN THE LITERATURE. ALL THE MODELS WERE TRAINED ON THE SMDD DATASET, AND EVALUATED ON THE DATASET SPECIFIED ON THE LEFT COLUMN OF THE TABLE. ALL THE RESULTS ARE IN PERCENTAGE (%) AND THE BEST ARE IN BOLD.

			BPCER @ APCER =	
Test	Model	EER	1%	20%
FRLL-Style-GAN2	Inception	11.37	72.06	6.86
	PW-MAD	16.64	80.39	13.24
	MixFacenet	8.99	42.16	4.41
	OrthoMAD	6.54	13.74	3.76
	IDistill (Ours)	**1.96**	**8.51**	**0.08**
FRLL-WebMorph	Inception	9.86	53.92	2.94
	PW-MAD	16.65	80.39	13.24
	MixFacenet	12.35	80.39	7.84
	OrthoMAD	15.23	70.92	9.50
	IDistill (Ours)	**4.01**	**14.41**	**0.33**
FRLL-OpenCV	Inception	5.38	38.73	0.98
	PW-MAD	2.42	22.06	0.49
	MixFacenet	4.39	26.47	1.47
	OrthoMAD	**0.73**	**0.73**	0.32
	IDistill (Ours)	2.46	6.14	**0.16**
FRLL-AMSL	Inception	10.79	72.06	4.90
	PW-MAD	15.18	96.57	5.88
	MixFacenet	15.18	49.51	11.76
	OrthoMAD	14.80	65.05	10.89
	IDistill (Ours)	**4.00**	**21.10**	**2.85**
FRLL-FaceMorpher	Inception	3.17	30.39	0.49
	PW-MAD	2.20	26.47	**0.00**
	MixFacenet	3.87	23.53	0.49
	OrthoMAD	**0.98**	**2.37**	0.08
	IDistill (Ours)	2.05	4.26	0.16

that IDistill is fairly more consistent, as such, the improvements on the databases where it surpasses the literature are much wider than the loss in the performance on the two other databases. Looking at the most extreme examples, in FRLL-OpenCV the EER of our architecture is only 1.73 percentual points larger than the value obtained by OrthoMAD, while IDistill decreases the EER in 11.22 percentual points when tested in the FRLL-WebMorph dataset, which constitutes a much more relevant difference in performance. While looking beyond EER it is also possible to see a wide improvement on the BPCER@APCER at both 1% and 20%. Moreover, on FRLL-OpenCV the higher EER of IDistill is mitigated by a lower BPCER@APCER = 20%.

When compared to OrthoMAD, our method presents an architecture with the same computation cost on inference, but significantly more interpretable, since OrthoMAD does not guarantee that the information yield by both vectors is related to identity. We are capable of identifying not only attacks, but justify utilising the information of which vectors contain the identity, and which do not. Due to the approximation between the autoencoder latent space and the IDistill latent space, it might also be possible to reconstruct parts of the identity utilising the decoder. While not used in this study, the information regarding the intensity of the vectors extracted by the morphing classifier, v_1 and v_2 might also allow to infer the morphing percentage associated with each fused identity, which might be useful in future works.

VI. CONCLUSION

In this document we have presented a novel method for face morphing attack detection that is interpretable, compact and performs at the state-of-the-art level. The proposed IDistill method was trained utilising a two step scheme based on the training of an autoencoder to reconstruct *bonafide* images, and a distillation step integrated on the standard training of a morphing classifier, utilizing the encoder as teacher and the first part of the classifier as student.

While we relaxed the orthogonality constraint from previous methods, we devised a more consistent and reliable solution to ensure that the identity information is, in fact, separated in two individual vectors. Moreover, we dismiss any concatenation of these vectors, ensuring an interpretable analysis of the scores produced by each and their contribution to the final prediction. As future work on the interpretability capabilities of this study, it would be interesting to explore the reconstruction capabilities utilising the identity vectors and the decoder model from the autoencoder architecture. Another possible direction is to verify whether the intensities of the vectors extracted by the morphing classifier allow to quantify the morphing percentage of the identities that were fused to generate each attack sample.

Overall, IDistill surpasses the performance of the previous methods published in the literature, while ensuring the advantages previously mentioned. In some scenarios the performance is drastically better. There is much work to be done on

the topic of face morphing attack detection, nonetheless, IDistill is a step forward towards the integration of interpretable approaches that are competitive with fully black-box systems.

ACKNOWLEDGMENT

This work is co-financed by Component 5 - Capitalization and Business Innovation, integrated in the Resilience Dimension of the Recovery and Resilience Plan within the scope of the Recovery and Resilience Mechanism (MRR) of the European Union (EU), framed in the Next Generation EU, for the period 2021 - 2026, within project NewSpacePortugal, with reference 11. It was also financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within the PhD grants "2020.06434.BD" and "2021.06872.BD". The research work has been also funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

REFERENCES

[1] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[2] P. C. Neto, J. R. Pinto, F. Boutros, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Beyond masks: On the generalization of masked face recognition models to occluded face recognition," *IEEE Access*, vol. 10, pp. 86 222–86 233, 2022.

[3] P. C. Neto, F. Boutros, J. R. Pinto, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Focusface: Multi-task contrastive learning for masked face recognition," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–08.

[4] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, "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)*. IEEE, 2018, pp. 1–10.

[5] I. Medvedev, F. Shadmand, and N. Gonçalves, "Mordeephy: Face morphing detection via fused classification," *arXiv preprint arXiv:2208.03110*, 2022.

[6] P. C. Neto, A. F. Sequeira, and J. S. Cardoso, "Myope models-are face presentation attack detection models short-sighted?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 390–399.

[7] P. C. Neto, A. F. Sequeira, J. S. Cardoso, and P. Terhörst, "Pic-score: Probabilistic interpretable comparison score for optimal matching confidence in single-and multi-biometric (face) recognition," *arXiv preprint arXiv:2211.12483*, 2022.

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

[9] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper, "Pw-mad: Pixel-wise supervision for generalized face morphing attack detection," in *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part I*. Springer, 2021, pp. 291–304.

[10] M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso *et al.*, "Synmad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–10.

[11] P. C. Neto, T. Gonçalves, M. Huber, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Orthomad: Morphing attack detection through orthogonal identity disentanglement," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2022, pp. 1–5.

[12] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros, "Privacy-friendly synthetic data for the development of face morphing attack detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1606–1617.

[13] N. Damer, M. Fang, P. Siebke, J. N. Kolf, M. Huber, and F. Boutros, "Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders," *arXiv preprint arXiv:2302.01843*, 2023.

[14] M. Fang, F. Boutros, and N. Damer, "Unsupervised face morphing attack detection via self-paced anomaly detection," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–11.

[15] S. Banerjee, P. Jaiswal, and A. Ross, "Facial de-morphing: Extracting component faces from a single morph," in *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2022, pp. 1–10.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

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

[18] L. DeBruine and B. Jones, "Face research lab london set," Apr 2021. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666

[19] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," *arXiv preprint arXiv:2012.05344*, 2020.

[20] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 104–12 114, 2020.

[21] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Can gan generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2020, pp. 1–6.

[22] L. DeBruine, "debruine/webmorph: Beta release 2," *Zenodo https://doi.org/10*, vol. 5281, 2018.

[23] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, no. 4, pp. 325–332, 2018.

[24] A. Quek, "Facemorpher," 2019.

[25] S. Mallick, "Face morph using opencv — c++ / python — learnopencv," Mar 2016. [Online]. Available: https://learnopencv.com/face-morph-using-opencv-cpp-python/

[26] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "Cr-fiqa: face image quality assessment by learning sample relative classifiability," *arXiv preprint arXiv:2112.06592*, 2021.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[29] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper, "Mixfacenets: Extremely efficient face recognition networks," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.