OrthoMAD: Morphing Attack Detection Through Orthogonal Identity Disentanglement

Pedro C. Neto[1,2], Tiago Gonçalves[1,2], Marco Huber[3,4], Naser Damer[3,4],
Ana F. Sequeira[1], and Jaime S. Cardoso[1,2]
[1]*Centre for Telecommunications and Multimedia, INESC TEC, Porto, Portugal*
[2]*Faculdade de Engenharia da Universidade do Porto, Porto, Portugal*
[3]*Fraunhofer Institute for Computer Graphics Research (IGD), Darmstadt, Germany*
[4]*Technische Universität Darmstadt, Darmstadt, Germany*
pedro.d.carneiro@inesctec.pt

2022 International Conference of the Biometrics Special Interest Group (BIOSIG) | 978-1-6654-7666-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIOSIG55365.2022.9897057

***Abstract*—Morphing attacks are one of the many threats that are constantly affecting deep face recognition systems. It consists of selecting two faces from different individuals and fusing them into a final image that contains the identity information of both. In this work, we propose a novel regularisation term that takes into account the existent identity information in both and promotes the creation of two orthogonal latent vectors. We evaluate our proposed method (OrthoMAD) in five different types of morphing in the FRLL dataset and evaluate the performance of our model when trained on five distinct datasets. With a small ResNet-18 as the backbone, we achieve state-of-the-art results in the majority of the experiments, and competitive results in the others.**

***Index Terms*—Face, Presentation Attack Detection, Morphing, Identity Disentanglement, ResNet-18**

I. Introduction

Over the years, the performance of face recognition systems kept increasing significantly. Larger datasets have been collected, and better models have been developed and published [1]. Hence, its usage has been proportional to its performance, leading to a wide spread of these algorithms. However, in the real world, these models might face attacks that aim to increase the number of false positives given by the model [2]. These attacks vary significantly. For instance, one may add a simple facial mask to a user's face or print the target user's face and use it as a presentation attack to the system. Hence, researchers also focus on developing detection approaches to these threats [3], [4]. The former does not care about which identity the model identifies, whereas the latter aims to create a positive sample based on the presented identity.

Another category of attacks, known as morphing attacks, aims to incorporate facial features from two distinct identities in a single image crafted from a fusion of the images corresponding to these two identities [5]. Hence, one may use this newly designed image to allow two distinct users to enter the system or two different people to use the same passport and pass border control. Besides, similarly to the research on morphing attacks, the community has also dedicated several research efforts toward detecting these attacks [6]–[9]. Usually, these methods do not use any information regarding the identity fused in the attacks and focus the model's training on detecting a potential attack. Therefore, it is often common to formulate these problems as binary classification tasks where the negative label is associated with an attack and the positive with a *bona fide* sample.

The work proposed in this paper includes a new regularisation term to a modified ResNet-18 [10] architecture trained to detect morphing attacks. Besides the binary classification system, this new framework presents two vectors in different latent spaces. These vectors, known as identity vectors, are strongly regularised to be orthogonal, thus, promoting independent identity information in each. Moreover, although we used ResNet-18 in our experiments, one may apply this methodology on top of any existing architecture.

The main contributions of this work are: 1) the addition of a loss regularisation term that aims to promote orthogonality between the identity-related embeddings (i.e. latent space vectors); 2) the empirical validation of the proposed loss in a large set of publicly available datasets of face morphing attacks; 3) a comparison with the state-of-the-art approaches presented in the literature that use the same dataset.

In this Introduction, we discriminated the types of attacks that affect biometric systems and which we are trying to detect and enumerated our main contributions. Moreover, we organised the remainder of this paper as follows: Section II describes in detail the datasets used to design the experiments; Section III presents the architecture design of the proposed regularisation approach; Section IV reports and discuss the performance of our method in the selected datasets; and Section V, concludes this paper with some final thoughts and future work suggestions. The code related to this paper is publicly available in a GitHub repository[1].

This work was financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project LA/P/0063/2020 and the PhD grants "2021.06872.BD" and "UIDB/50014/2020". This 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.

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

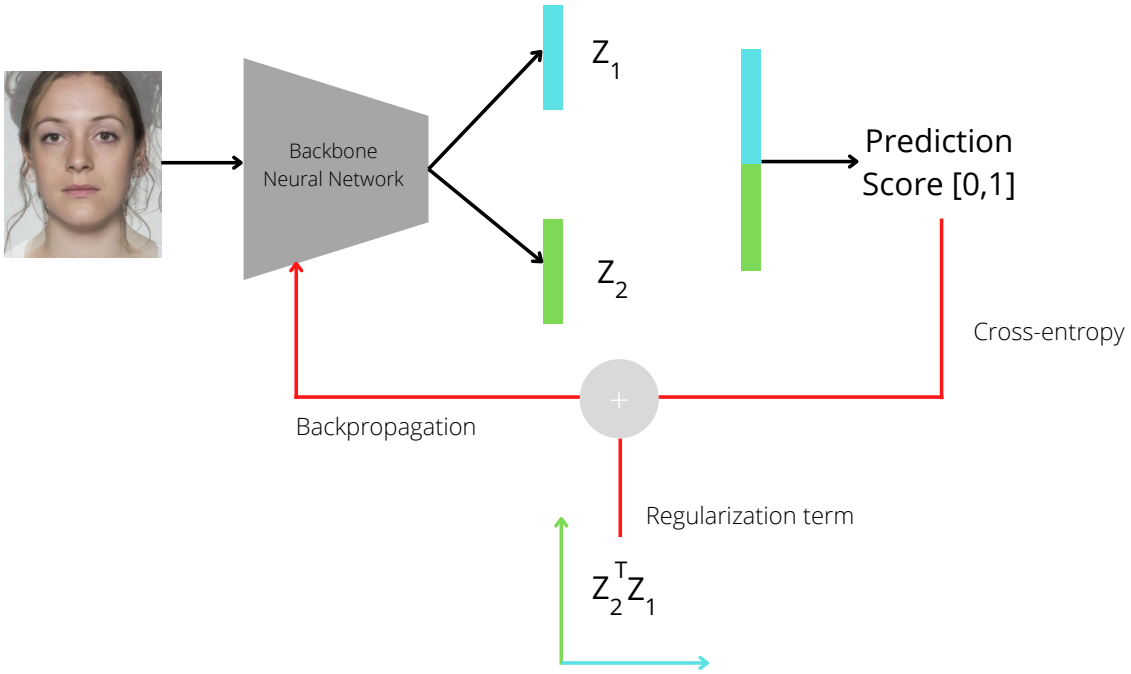

Fig. 1: Overview of the architecture design for the integration of the new regularisation term. The regularisation loss is applied at the vector (Z_1 and Z_2), hence, their gradients do not affect the final classification layer. The system is composed of three components, a backbone network, two identity vectors and a final classification layer that uses the concatenation of both vectors.

II. Databases

a) MorGAN: Divided into two sets, the MorGAN dataset [6] fused images from the CelebA [11] dataset as morphing attacks. Images were selected based on their frontal pose, and the OpenFace [12] FR solution was used to find the most similar pairs to morph [13]. The morphing approach followed relied on the landmark-based OpenCV [14] or a GAN-based [6] solution. These different approaches resulted in the two distinct sets MorGAN-LMA and MorGAN-GAN, respectively. Each of the train and test sets (available in both sets) contains 500 morphing attacks and 750 *bona fide* samples. This dataset contains images of low resolution (64x64).

b) FRLL: Extensively used to evaluate morphing attack detection algorithms, the FRLL-Morphs dataset [15] was created from the publicly available Face Research London Lab dataset [16]. The dataset contains five distinct morphing approaches, including Style-GAN2 [17], [18], WebMorph [19], AMSL [20], FaceMorpher [21] and OpenCV [14]. Each of the five approaches contains 1222 morphed faces created from frontal faces with high resolution and 204 *bona fide* samples. This database does not contain disjoint train/test sets, hence it was used just for evaluation.

c) LMA-DRD: The LMA-DRD dataset [7] uses the images published with the VGGFace2 [22] dataset to create the morphing attack. Despite having more than 3 million images available, not all were selected to be fused into attacks. The selection requirements included a frontal pose, high resolution, and a neutral expression. Using the parametrisation introduced in [23], the images were morphed with the OpenCV morphing [14]. *Bona fide* images were selected using the previously mentioned criteria. This dataset has two versions, a digital - LMA-DRD (D) - and a printed and scanned - LMA-DRD (PS). We used the train set for training (96 morphs and 121 BF images), and the identity-disjoint test set for testing (88 morphs and 123 *bona fide* (BF) images).

d) SMDD: The Synthetic Morphing Attack Detection Development (SMDD) [24] utilised the official open-source implementation of StyleGan2-ADA [17] to generate 500k images of faces using a random Gaussian noise vector sampled from a normal distribution. From these images, 50k were selected due to their high quality (using CR-FIQA [25]), and 25k from those were considered *bona fide*. 5k of the non-*bona fide* images were selected as key morphing images and paired with five randomly chosen non-*bona fide* images. They were then morphed with the OpenCV [14] approach, generating 15k attack samples.

III. Methodology

Morphing attacks result from a fusion process of two distinct identities. In the final image, there is enough information about both identities to trick a face recognition system. Hence, to leverage the presence of such information, we designed a regularisation term to promote the separation

Tab. I: Results for the different types of morphing techniques included in the FRLL dataset. For morphing type, we present the performance of the model per training dataset. All the results are in percentage (%), and the best are in bold.

				BPCER @ APCER =	
Test	Train	Model	EER	1%	20%
FRLL - Style-GAN2	MorGAN-LMA	ResNet-18	**28.72**	99.01	**60.81**
		OrthoMAD (Ours)	51.47	**97.46**	81.51
	MorGAN-GAN	ResNet-18	53.60	99.59	87.07
		OrthoMAD (Ours)	**17.34**	**59.81**	**14.32**
	LMA-DRD (D)	ResNet-18	62.60	100.00	97.87
		OrthoMAD (Ours)	**31.50**	**99.91**	**63.74**
	LMA-DRD (PS)	ResNet-18	71.19	100.00	97.38
		OrthoMAD (Ours)	**29.21**	**99.59**	**53.02**
	SMDD	ResNet-18	11.62	35.18	7.77
		OrthoMAD (Ours)	**6.54**	**13.74**	**3.76**
FRLL - WebMorph	MorGAN-LMA	ResNet-18	37.10	99.09	60.68
		OrthoMAD (Ours)	**8.35**	**45.61**	**2.53**
	MorGAN-GAN	ResNet-18	23.25	89.59	25.71
		OrthoMAD (Ours)	**14.41**	**45.12**	**9.90**
	LMA-DRD (D)	ResNet-18	93.85	100.00	100.00
		OrthoMAD (Ours)	**23.58**	**97.37**	**29.40**
	LMA-DRD (PS)	ResNet-18	88.62	100.00	100.00
		OrthoMAD (Ours)	**10.65**	**60.28**	**7.20**
	SMDD	ResNet-18	16.29	85.33	16.29
		OrthoMAD (Ours)	**15.23**	**70.92**	**9.50**
FRLL - AMSL	MorGAN-LMA	ResNet-18	54.16	99.63	84.36
		OrthoMAD (Ours)	**0.91**	**0.91**	**0.00**
	MorGAN-GAN	ResNet-18	44.59	**90.94**	63.72
		OrthoMAD (Ours)	**7.91**	97.01	**1.19**
	LMA-DRD (D)	ResNet-18	50.67	95.86	75.86
		OrthoMAD (Ours)	**30.43**	**89.70**	40.46
	LMA-DRD (PS)	ResNet-18	83.35	100.00	100.00
		OrthoMAD (Ours)	**27.26**	**87.21**	**37.19**
	SMDD	ResNet-18	34.43	65.97	27.26
		OrthoMAD (Ours)	**14.80**	**65.05**	**10.89**
FRLL - FaceMorpher	MorGAN-LMA	ResNet-18	**5.48**	**25.45**	**1.23**
		OrthoMAD (Ours)	34.20	97.29	54.90
	MorGAN-GAN	ResNet-18	37.15	97.95	**55.72**
		OrthoMAD (Ours)	**35.27**	**94.10**	56.71
	LMA-DRD (D)	ResNet-18	39.60	99.59	76.51
		OrthoMAD (Ours)	**30.19**	**83.87**	**38.13**
	LMA-DRD (PS)	ResNet-18	40.02	**99.42**	**70.21**
		OrthoMAD (Ours)	**34.12**	99.83	71.84
	SMDD	ResNet-18	2.95	5.32	2.37
		OrthoMAD (Ours)	**0.98**	**2.37**	**0.08**
FRLL - OpenCV	MorGAN-LMA	ResNet-18	**16.29**	**64.61**	**11.62**
		OrthoMAD (Ours)	27.92	93.44	38.82
	MorGAN-GAN	ResNet-18	40.29	99.34	79.11
		OrthoMAD (Ours)	**29.07**	**97.21**	**43.48**
	LMA-DRD (D)	ResNet-18	66.75	99.83	96.56
		OrthoMAD (Ours)	**33.98**	**99.59**	**60.68**
	LMA-DRD (PS)	ResNet-18	49.63	100.00	85.17
		OrthoMAD (Ours)	**42.17**	**99.09**	**73.87**
	SMDD	ResNet-18	1.22	11.8	0.41
		OrthoMAD (Ours)	**0.73**	**0.73**	**0.32**

of the information from both identities. When two vectors are orthogonal, there is no common information in them. The orthogonality of these two vectors is then the desired property when trying to disentangle identity information into separate vectors. And thus, we introduce a regularisation term (Eq. 1), which leverages the inner product of two vectors.

$$Reg = (Z_1^T Z_2)^2 \tag{1}$$

Its integration in the final loss is straightforward since the

Tab. II: Results comparison with three other models published in the literature. All the models were trained on the SMDD dataset. 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 (Ours)	**6.54**	**13.74**	**3.76**
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 (Ours)	15.23	70.92	9.50
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 (Ours)	**0.73**	**0.73**	**0.32**
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 (Ours)	14.80	65.05	10.89
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 (Ours)	**0.98**	**2.37**	0.08

main objective is to minimise the inner product to zero. We further square the inner product to make the optimisation smoother. To be able to have these two vectors, the architecture of a ResNet-18 had to be adapted. The last fully-connected layer was replaced with two fully-connected layers that output a vector of size 32 each. These vectors are used to apply the orthogonal regularisation and are afterwards concatenated (Eq. 2).

$$Z = concat(Z_1, Z_2) \tag{2}$$

As seen in Fig. 1, this concatenation is directly fed into another fully connected layer to produce a score Y, which is activated with the sigmoid non-linearity (Eq. 3).

$$Y = \frac{1}{1 + e^{-(W^T Z)}} \tag{3}$$

The loss of the prediction is calculated using the Binary Cross-Entropy (Eq. 4), and the regularisation term is summed to the final loss with a certain weight controlled by the hyperparameter α (Eq. 5). An illustration of the end-to-end optimisation and classification process is seen in Fig. 1.

$$\textit{Binary Cross Entropy} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{4}$$

$$Loss = -(y \log(p) + (1 - y) \log(1 - p)) + \alpha (Z_2^T Z_1)^2 \tag{5}$$

a) Experimental setup: We design the experiments to evaluate the performance of our approach when compared to a baseline architecture that does not include regularisation. For this, we train both models in all the datasets and evaluate them in the five different types of morphing approaches found in the FRLL dataset. We trained all the models with a batch size of 16, a learning rate of 10^{-5}, and an α hyperparameter of 100. We further augmented the dataset with random horizontal flips and resized all the images to be 224×224. Finally, we also crop the faces from the original images. The detection performance is shown by the Attack Presentation Classification Error Rate (APCER) (i.e., attacks classified as *bona fide*), and the *Bona fide* Presentation Classification Error Rate (BPCER) (i.e., the *bona fide* samples classified as attacks). We report the BPCER at two different fixed APCER values (1.0% and 20.0%). We also report the equal error rate (EER), which is the BPCER and APCER at the decision thresholds where they are the same.

IV. Results and Discussion

In this Section, we discuss the results obtained with our regularisation term (OrthoMAD) and compare it with a version without the regularisation. The performance of OrthoMAD indicates an EER lower than the EER for ResNet-18 in 22 out of 25 evaluations (see Tab. I). Nonetheless, the values of BPCER @ APCER $\{1\%,20\%\}$ seem to vary more when the EER is close in both models. While the performance of OrthoMAD indicates good performance and generalisation capabilities across datasets, it is worth noting that it has some difficulties when handling certain types of morphing (e.g., FaceMorpher) when the training dataset is small. Hence, this model excels in the majority of the testing datasets when trained in a large-scale training set (SMDD), even when the dataset is solely composed of synthetic images, which might hold less "identity" information.

We further extend our comparison to include the models published in the literature (see Tab. II). Hence, we compare with the results shown by Damer *et al.* [24] on the Inception [26], the PW-MAD [7] and the MixFacenet [27] models. OrthoMAD outperforms all these methods in three out of

five evaluation settings and is competitive in the remaining two. These results show that even with a less parameterised backbone network, our regularisation term was capable of achieving state-of-the-art performance in the task of morphing attack detection.

V. Conclusions and Future Work

This paper presented a novel face-morphing attack detection system that promotes the disentanglement of identity features. We proposed to enforce orthogonality between the learned features of two identities in an input image, using a novel term in the loss function (i.e., regularisation). We compared our model against the state-of-the-art and obtained similar results. Nevertheless, we argue that this novel term in the loss function imposes a constraint that contributes toward the transparency of our model. With this regularisation method, we are sure that the user knows *a priori* one of the rules the model must follow to output a decision. Further work should be devoted to testing different backbone architectures (e.g., attention-based models) and to the generation of saliency map-based explanations to assess if the imposition of our constraints has some significant impact on post-hoc explanation methods.

References

[1] P. J. Grother, M. L. Ngan, K. K. Hanaoka *et al.*, "Ongoing face recognition vendor test (frvt) part 2: Identification," 2018.

[2] R. S. Kramer, M. O. Mireku, T. R. Flack, and K. L. Ritchie, "Face morphing attacks: Investigating detection with humans and computers," *Cognitive research: principles and implications*, vol. 4, no. 1, pp. 1–15, 2019.

[3] 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.

[4] P. C. Neto, F. Boutros, J. R. Pinto, N. Darner, 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.

[5] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[6] 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.

[7] 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 *International Symposium on Visual Computing*. Springer, 2021, pp. 291–304.

[8] J. E. Tapia and C. Busch, "Single morphing attack detection using feature selection and visualization based on mutual information," *IEEE Access*, vol. 9, pp. 167 628–167 641, 2021.

[9] M. Huber, F. Boutros, A. T. Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso, J. Tremoço, M. Lourenço, S. Serra, E. Cermeño, M. Ivanovska, B. Batagelj, A. Kronovšek, P. Peer, and V. Štruc, "Syn-mad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data." arXiv, 2022. [Online]. Available: https://arxiv.org/abs/2208.07337

[10] 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.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

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

[15] 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.

[16] 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

[17] 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.

[18] 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.

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

[20] 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.

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

[22] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[23] R. Raghavendra, K. Raja, S. Venkatesh, and C. Busch, "Face morphing versus face averaging: Vulnerability and detection," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 555–563.

[24] 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.

[25] 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.

[26] 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.

[27] 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.