Face Morphing Attack Detection Using Privacy-Aware Training Data

Marija Ivanovska[1], Andrej Kronovšek[2], Peter Peer[2], Vitomir Štruc[1], Borut Batagelj[2]

[1]*Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, SI-1000 Ljubljana, Slovenia*
[2]*Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, SI-1000 Ljubljana, Slovenia*
E-mail: marija.ivanovska@fe.uni-lj.si

Abstract

Images of morphed faces pose a serious threat to face recognition–based security systems, as they can be used to illegally verify the identity of multiple people with a single morphed image. Modern detection algorithms learn to identify such morphing attacks using authentic images of real individuals. This approach raises various privacy concerns and limits the amount of publicly available training data. In this paper, we explore the efficacy of detection algorithms that are trained only on faces of non–existing people and their respective morphs. To this end, two dedicated algorithms are trained with synthetic data and then evaluated on three real-world datasets, i.e.: FRLL-Morphs, FERET-Morphs and FRGC-Morphs. Our results show that synthetic facial images can be successfully employed for the training process of the detection algorithms and generalize well to real-world scenarios.

1 Introduction

Nowadays, the vast majority of applications for person identity verification rely on Face Recognition Systems (FRSs), which match a human face to an entry from a database of faces. Modern FRSs have proven to be highly accurate when genuine faces are presented to the system [7]. They are however prone to various attacks, whose aim is to gain illegal access by false authentication [8].

Lately, face morphs have become a growing concern for the reliability of face verification systems. A face morph is a composite image generated from two (or more) facial images of distinct subjects. Recent advances in generative deep models have enabled an almost effortless generation of realistic and high-quality morphed facial images. Such images can be utilized to verify all identities that have been used in the morph-generation process. A successful detection of *face morphing attacks* is therefore critical for the prevention of illegal activities [4].

Various *Morphing Attack Detection* (MAD) algorithms have been proposed over the years to automatically distinguish real from morphed faces. However, regardless of the detection technique, the training of these models requires a massive database of genuine face images. Training protocols therefore raise various privacy-related concerns and limit the amount of publicly available training data that can be used to learn MAD models. In this paper, we address the privacy issues related to MADs by exploring the idea of using synthetic training data, as illustrated in Figure 1. For this purpose, we use the SMDD [3] dataset, where StyleGAN2 [10], a state-of-the-art Generative Adversarial Network (GAN), was utilized to generate *bona fide* face images of non-existing people. These images were then used for the generation of the face morphs. With this dataset, we train two powerful binary classifiers, Xception and HRNet, and evaluate their detection performance on three real-world datasets – FRLL-Morphs, FERET-Morphs and FRGC-Morphs. The results of the evaluation show that well-performing MAD models can be learned from synthetic data alone, and that the model generalize well over three diverse real-world morph datasets.

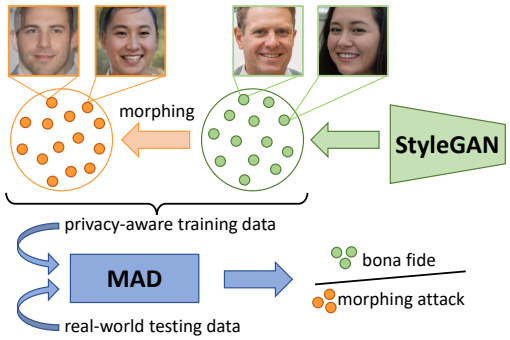

Figure 1: To avoid privacy related concerns in the development of morphing attack detectors (MAD), we explore the idea of using synthetic training faces of non–existing people. Trained MADs are then tested on real–world datasets.

2 Related work

Existing morphing-attack-detection models can in general be categorized as single–image (S-MAD) or differential (D-MAD) MADs, depending on whether the face morph is examined independently or is compared to a reference sample. While D-MADs can be very accurate in closed–set problems, S-MADs aim to detect attacks without any prior knowledge about human identities. In this section, we only review S-MADs, since they are more closely related to our work.

Regardless of the face morphing technique used, the

Supported in parts by the ARRS project J2-1734 (B), and the ARRS research programmes P2-0250 (B) and P2-0214 (B).

generated morphs usually contain image irregularities such as artifacts, noise, pixel discontinuity, distortions, spectrum discrepancies, inconsistent illumination, etc. In the past, shallow algorithms, that implement extraction of photo-response non-uniformity (PRNU) noise [20] or reflection analysis [21] have been successfully employed for the detection of morphing attacks. Some other hand–crafted MAD methods have also used texture–based descriptors, such as LBP [13], LPQ [14] or SURF [11]. Although these methods achieved promising results, they were shown to have limited generalization capabilities. Moreover, as the face morphing techniques improved over time, the performance of shallow methods became less competitive, as they struggled to detect modern, deep-learning generated or heavily post–processed face morphs.

More recent MAD models take advantage of the development of data–driven, deep-learning algorithms. Raghavendra *et al.* [17] were amongst the first to propose transfer learning. In their work, attacks are detected with a simple, fully–connected binary classifier, fed with fused VGG19 and AlexNet features, pretrained on ImageNet. Wandzik *et al.* [23], on the other hand, achieve highly accurate results with features from general–purpose face recognition systems (FRSs) combined with an SVM. Ramachandra *et al.* [18] utilize Inception in a similar manner, while Damer *et al.* [4] argue, that pixel–wise supervision, where each pixel is classified as a bona fide or a morphing attack, is superior, when used in addition to the binary, image–level objective. Recently, MixFaceNet [1] by Boutros *et al.* achieved state–of–the–art results in different face–related detection tasks, including face morphing detection [3]. This model represents a highly efficient architecture that captures different levels of face attack cues by using differently sized convolutional kernels.

3 Methods

We consider two different classification models, Xception and HRNet, to detect face morphing attacks in this study and train them using synthetic data only. The two models represent the entries from the University of Ljubljana to the recent *Face Morphing Attack Detection Competition based on Privacy-aware Synthetic Training Data* (SYN-MAD), held in conjunction with the 2022 International Joint Conference on Biometrics (IJCB 2022) [8], which achieved the best and third best overall performance among all submitted entries.

Xception [2] is a convolutional neural network (CNN) that updates and simplifies the architecture of the InceptionV3 model [22] by replacing the Inception modules with depth–wise separable convolutions. We use Xception as a feature extractor, while the binary classification is performed by a fully connected two–layer network. The output layer consist of 2 neurons, followed by a softmax activation function. Similar to previous research, we use cross–entropy as the learning objective.

HRNet [24] is again a CNN that unlike other networks maintains high–resolution representations of the input sample through the whole feedforward process. Such an architecture contributes to more descriptive image representations, which was proven to improve the results

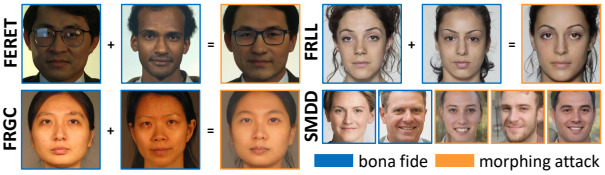

Figure 2: Examples of bona fide and morphing attack images from the SMDD [3] training dataset and the testing datasets FERET-Morphs, FRLL-Morphs and FRGC-Morphs [19].

Table 1: Number of bona fide images (BF), number of morphing attacks (MA) generated by morphing methods OpenCV (OCV), FaceMorpher (FM), StyleGAN (SG), AMSL, Webmorpf (WM) and image size of samples in each dataset.

Dataset	Image size		BF	OCV	FM	SG	AMSL	WM
FRLL-M	1350	1350	204	1221	1222	1222	2175	1221
FERET-M	512	768	1:413	529	529	529	/	/
FRGC-M	227	277	3:167	964	964	964	/	/

of different classification tasks. In our experiments, we replace the classification head of HRNet with a two–layer classification module, to perform binary detection of bona fide images and morphing attacks. The output layer consists of only one neuron, followed by a sigmoid activation function. In the training phase, the parameters are optimized using the binary cross–entropy loss.

4 Experiments

4.1 Datasets

We use one synthetic and three publicly available real–world datasets in this work. The training is done exclusively with the synthetic data, while the evaluation is performed on three commonly used face morphing datasets.

Training data. For training, we use the SMDD dataset [3], provided by the organizers of the SYN-MAD competition [8]. The dataset consists of 25;000 bona fide and 15;000 morphed images of size 256 256 pixels. Bona fide instances represent carefully selected images from a set of randomly generated StyleGAN2 [10, 9] faces. A separate, non – overlapping StyleGAN2 image set was used for the generation of face morphing attacks. Face morphs were created using the landmark–based morphing technique from OpenCV[1]. A few selected samples from the SMDD dataset are presented in Figure 2.

Testing data. The trained MAD models are tested on three diverse morphing datasets proposed by Sarkar *et al.* in [19], i.e. FRLL- Morphs, FERET-Morphs and FRGC-Morphs. All face morphs were created by combining bona fide images from their respective face datasets, i.e. FRLL [5, 12], FERET [16] and FRGC [15]. To generate landmark–based morphs, the authors used OpenCV and FaceMorpher[2], while deep-learning–based morphs are generated with StyleGAN2. In addition to these methods, AMSL[12] and Webmorph[3] are also used, but only for the images from the FRLL dataset. Information about image sizes and the number of samples per morphing

[1] https://learnopencv.com/face-morph-using-opencv-cpp-python/
[2] https://github.com/alyssaq/face_morpher
[3] https://webmorph.org/

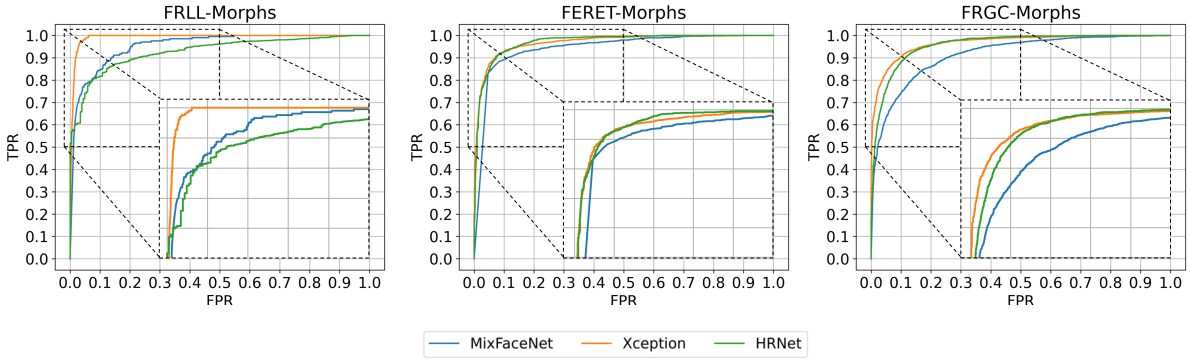

Figure 3: ROC curves generated on FRLL-Morphs, FERET-Morphs and FRGC-Morphs for the tested models. Note that HRNet achieves very competitive results on FERET-Morphs and FRGC-Morphs, but performs the weakest on FRLL-Morphs. Xception on the other hand, consistently outperforms the baseline method MixFaceNet, on all considered datasets.

method is given in Table 1. Selected samples from all three datasets are presented in Figure 2.

4.2 Experimental setup

In our experiments, we first preprocess images from all datasets by cropping out the facial areas. Bounding boxes of the SMDD are provided by the authors of the dataset. For the other three databases, we use RetinaFace [6], to localize the facial region-of-interest. Prior to their use, cropped images are resized to 299 299 pixels for Xception and 256 256 pixels for HRNet. Additionally, the training data is augmented with horizontal flips to increase the amount of data available and avoid overfitting.

The CNNs were optimized using the Adam optimizer, with a learning rate of 0:0001. The models were trained from scratch for 30 full epochs, with a batch size of 16. After each training epoch, the classification accuracy of the networks was calculated on a small holdout set of each test dataset. The best performing parameters on each of the three datasets were saved as the final model for that particular dataset. The code was implemented in Python 3.8 with PyTorch 1.9 and CUDA 11.6. Experiments were run on a single GeForce GTX 1080 Ti. The computational complexity of Xception is 11 GFLOPs, while HRNet has 34 GFLOPs.

5 Results

In Figure 3 and Table 2 we present the results obtained with our two models, Xception and HRNet, and the baseline MixFaceNet-MAD from [8]. The weights of MixFaceNet-MAD, optimized on the SMDD dataset, were provided by the authors of the model. In our experiments, the best overall results were achieved by Xception, whose Equal Error Rates (EER) are 3:26%, 8:25% and 9:75% on FRRL–Morphs, FERET-Morphs and FRGC–Morphs, respectively (Table 2). The runner-up, HRNet, achieves a similar performance on FERET-Morphs and FRGC-Morphs. However, among the tested models, HRNet is least successful on FRLL-Morphs, where it achieves an EER of 13:73%. On this dataset, MixFaceNet yields a slightly better performance than HRNet with an EER of 12:18%, but is outperformed by both, Xception and HRNet, on the other two databases, i.e. FERET-Morphs and FRGC-Morphs. The complete ROC curves of the experiments

Table 2: Detection results for MAD methods MixFaceNet (MFN), Xception (XN) and HRNet (HRN) on the real–world datasets FRLL-Morphs (FRLL-M), FERET-Morphs (FERET-M) and FRGC-Morphs (FRGC-M). Best scores per dataset are marked blue, while runner–up results are marked orange. All three models were trained on the synthetic SMDD dataset [3].

MAD	Test data	AUC(%)	EER(%)	BPCER (%) @ APCER =			
				0:10%	1:00%	10:00%	20:00%
MFN [3]	FRLL-M	95:43	12:18	100:0	100:0	15:20	5:88
	FERET-M	94:27	10:65	100:0	100:0	11:75	6:51
	FRGC-M	91:42	16:36	100:0	64:86	25:89	14:02
XN [2]	FRLL-M	**99:17**	**3:26**	**85:29**	**28:92**	**0:49**	**0:0**
	FERET-M	**96:84**	**8:25**	**79:62**	**43:31**	**7:29**	**4:03**
	FRGC-M	**96:63**	**9:75**	**58:19**	**35:11**	**9:44**	**4:23**
HRN [24]	FRLL-M	92:79	13:73	100:00	42:84	18:65	11:12
	FERET-M	**97:05**	8:49	91:43	54:00	7:44	**2:27**
	FRGC-M	95:77	10:89	82:26	55:50	12:14	4:46

are visualized in Figure 3 and show a similar picture as the discussed numerical results.

To better understand the differences between the evaluated models, we additionally assess their performance on only one face morphing technique at a time. As can be seen from Figure 4, Xception shows the greatest generalization capabilities, when it comes to detection of different face morphing methods. HRNet provides very competitive results on face morphs generated by OpenCV, FaceMorpher and StyleGAN. Nevertheless, AMSL and Webmorph attacks are too challenging for this model. We hypothesize, that this might be due to the structure of the training data. Synthetic face morphs from SMDD are generated using only one morphing method, i.e. OpenCV. With such training data, models are in general at higher risk of overfitting to one specific type of morph. Since HRNet has far more trainable parameters than Xception and MixFaceNet, it is also more prone to overfitting, when trained on smaller datasets like SMDD.

6 Conclusion

In this paper, we tackle the privacy issues associated with the datasets used for the development of face morphing detection algorithms. To address related privacy concerns, we explore the idea of using a training database with faces of non–existing people, generated by StyleGAN. Using this data, we train three different MAD models and evaluate their performance on three commonly used

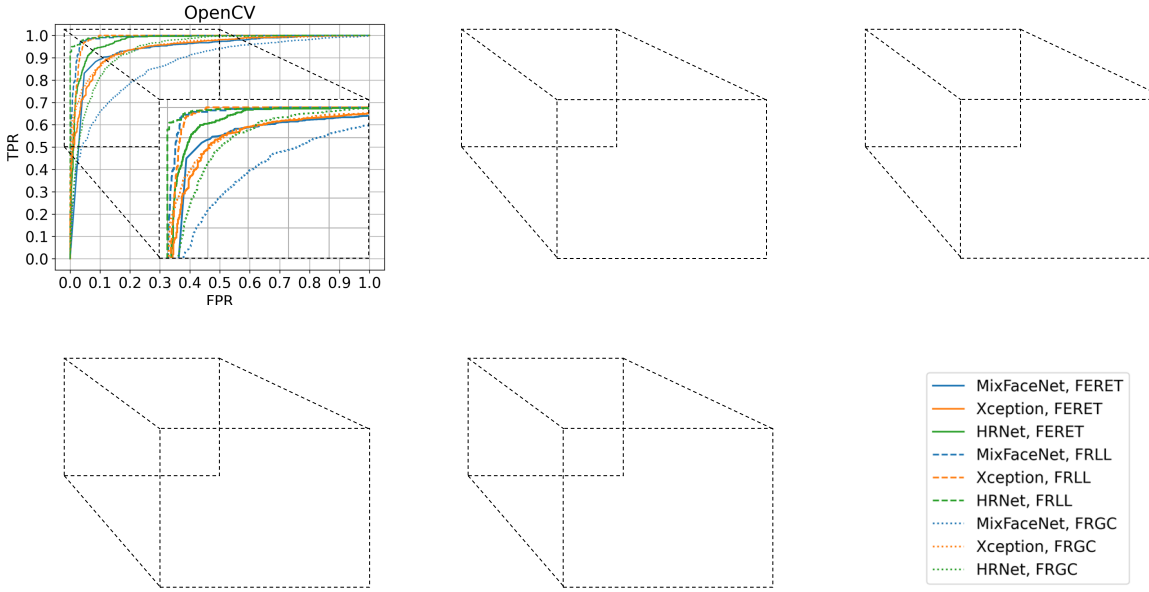

Figure 4: ROC curves generated for the tested MAD models for different types of face morphs. Note that the performance of the detectors differs quite considerably depending on the morphing procedure used.

real–world datasets. Our experiments show that in general, MAD models can be successfully trained on synthetic data and generalize well to real–world scenarios.

References

[1] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *IEEE IJCB*, pages 1–8, 2021.

[2] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *IEEE CVPR*, pages 1800–1807, 2017.

[3] 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. *IEEE CVPRW*, pages 1606–1617, 2022.

[4] 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*, pages 291–304. Springer International Publishing, 2021.

[5] L. DeBruine and B. Jones. Face Research Lab London Set, 2017.

[6] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *IEEE CVPR*, pages 5202–5211, 2020.

[7] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel. Strengths and weaknesses of deep learning models for face recognition against image degradations. *IET Biometrics*, 7(1):81–89, 2018.

[8] M. Huber, F. Boutros, A. Thi Luu, K. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Goncalves, A. F. Sequeira, J. S. Cardoso, T. Joao, M. Lourenc, S. Serra, E. Cermeno, M. Ivanovska, B. Batagelj, A. Kronovsek, P. Peer, and V. Struc. SYN-MAD 2022: Competition on Face Morphing Attack Detection Based on Privacy-aware Synthetic Training Data. In *IEEE IJCB*, 2022.

[9] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training Generative Adversarial Networks with Limited Data. In *NIPS*, 2020.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and Improving the Image Quality of StyleGAN. In *IEEE/CVF CVPR*, pages 8110–8119, 2020.

[11] A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, and P. Eisert. Dempster-Shafer Theory for Fusing Face Morphing Detectors. In *EUSIPCO*, pages 1–5, 2019.

[12] 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*, 7(4):325–332, 2018.

[13] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[14] V. Ojansivu and J. Heikkilä. Blur Insensitive Texture Classification Using Local Phase Quantization. In A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, editors, *Image and Signal Processing*, pages 236–243. Springer Berlin Heidelberg, 2008.

[15] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *IEEE CVPR*, volume 1, pages 947–954 vol. 1, 2005.

[16] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[17] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch. Transferable Deep-CNN Features for Detecting Digital and Print-Scanned Morphed Face Images. In *IEEE CVPRW*, pages 1822–1830, 2017.

[18] R. Ramachandra, S. Venkatesh, K. Raja, and C. Busch. Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features. In *CVIP*, pages 255–265, 2020.

[19] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Vulnerability Analysis of Face Morphing Attacks from Landmarks and Generative Adversarial Networks. 2020.

[20] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE TBBIS*, 1(4):302–317, 2019.

[21] C. Seibold, A. Hilsmann, and P. Eisert. Reflection Analysis for Face Morphing Attack Detection. In *EUSIPCO*, pages 1022–1026, 2018.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *CoRR*, 2015.

[23] L. Wandzik, G. Kaeding, and R. V. Garcia. Morphing Detection Using a General-Purpose Face Recognition System. In *EUSIPCO*, pages 1012–1016, 2018.

[24] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *TPAMI*, 2019.