

PW-MAD: Pixel-wise Supervision for Generalized Face Morphing Attack Detection

Naser Damer^{1,2}, Noémie Spiller¹, Meiling Fang^{1,2}, Fadi Boutros^{1,2},
Florian Kirchbuchner¹, and Arjan Kuijper^{1,2}

¹ Fraunhofer Institute for Computer Graphics Research IGD, Germany

² Department of Computer Science, TU Darmstadt, Germany

naser.damer@igd.fraunhofer.de

Abstract. A face morphing attack image can be verified to multiple identities, making this attack a major vulnerability to processes based on identity verification, such as border checks. Different methods have been proposed to detect face morphing attacks, however, with low generalizability to unexpected post-morphing processes. A major post-morphing process is the print and scan operation performed in many countries when issuing a passport or identity document. In this work, we address this generalization problem by adapting a pixel-wise supervision approach where we train a network to classify each pixel of the image into an attack or not during the training process, rather than only having one label for the whole image. Our pixel-wise morphing attack detection (PW-MAD) solution performs more accurately than a set of established baselines. More importantly, our approach shows high generalizability in comparison to related works, when evaluated on unknown re-digitized attacks. Additionally to our PW-MAD approach, we create a new face morphing attack dataset with digital and re-digitized attacks and bona fide samples, namely the LMA-DRD dataset that will be made publicly available for research purposes.

Keywords: Face recognition · Face morphing · Morphing attack detection.

1 Introduction

The face recognition performance advances driven by deep-learning [14], along with the relatively high social acceptance [3], have introduced face recognition technologies to security sensitive applications of identity management (e.g. ID/travel documents) [25]. However, face recognition systems are style vulnerable to many attacks [26,7], one of these is face morphing attacks. Ferrara et al. [16] analyzed the effect of face morphing attacks on face recognition early on by showing that one attack face image can be, automatically and by human experts, matched to more than one person. When morphing attacks are used in travel or identity documents, they allow multiple subjects to be verified to one document. This faulty subject link to the document identity can lead to a wide range of illegal activities, including financial transactions, illegal immigration, human trafficking, and circumventing criminal identity lists.

Morphing attack detection solutions (MAD) are developed to classify investigated face images into an attack or a bona fide (no attack). The performance

of MAD solutions has been shown to drop substantially when facing unknown (not used in the training) variations in the investigated images. Such variations include the morphing technique [41], image compression [24], and re-digitization (print and scan) [17] among others. The print of scan is one of the most studied of these variations as it represents a realistic scenario where a printed image is presented when applying for an ID/travel document. This image is then scanned to be included on the document and to go through a possible MAD. MAD solutions aiming at generalizability did utilise handcrafted [27,29] and deep learning features [17,29], however, while using single binary target per image in the training process.

This work aims at providing a generalizable MAD by proposing the adaption of pixel-wise supervision in the training process, giving the network the chance to distribute its focus more evident manipulation effects, rather than, general image non-attack related artifacts. This resulted in our pixel-wise supervised MAD (PW-MAD). To develop and evaluate our proposed solution, we additionally presented a publicly available morphing dataset that includes digital and re-digitized attacks and bona fide samples. Our solution proved to outperform a set of widely used baselines, especially when faced with unknown re-digitized images. Our PW-MAD has also shown better performance generalization over related works investigating the issue of re-digitized morphing attacks.

2 Related work

Morphing attacks have been first analyzed by Ferrera et al. [16], where they showed that morphed images can fool face recognition systems and human operators into matching the morphing attack image to multiple identities. From there, different methods have been proposed for MAD. These methods can be separated into two main categories, single image MAD and differential MAD [40]. Single image MAD only analyses the investigated image to build a decision of attack or bona fide (not attack) [28,29,11,1]. The differential MAD uses the investigated image and an image captured live (assuming that the realistic process allows for that). Differential MAD analyses the relation between both images or their respective representation to build a decision of attack or bona fide on the investigated image [5,35,10,37]. As the single image MAD demands fewer requirements on the process involving MAD, this work focuses on this MAD category.

From the single image MAD, solutions can be roughly categorized into ones using handcrafted features and ones using deep learning features. Such handcrafted features included Binarized Statistical Image Features (BSIF) [33,27], Local Binary Patterns (LBP) [9], Local Phase Quantization (LPQ) [28], or even features long established in the image forensic analyses such as the photo response non-uniformity (PRNU) [12]. The MAD solutions based on deep learning commonly used pre-trained networks with or without fine-tuning, such as different versions of VGG [17], AlexNet [29], or even networks trained for face recognition purposes such as OpenFace [11]. However, all these works, unlike this work, used a single binary label as the target of their training processes.

Many of these works have raised the issue of the generalizability of the MAD decision when facing variabilities in the face morphing or image handling process. Such variabilities included the unknown synthetic image generation processes [8,6,41], different original data sources [34], morphing pair selection protocols [10], image compression processes [24], and re-digitization (print and scan) [17,27,29]. All these variabilities have been shown to cause a drop in the MAD performance when they were unknown in the MAD training phase. The most practically relevant of these is the re-digitization, as it reflects the practice of requiring a printed image for travel/identity document issuance, where the authorities would scan this printed image. For this case, a number of private databases were created along with the analyses of the generalization of different MAD solutions such as different versions of VGGFace [17], AlexNet [17], the fusion of different pre-trained networks [29], and the BSIF handcrafted features along with an SVM classifier [27,29]. All these works have shown a substantial drop in the MAD performance on the re-digitized attacks when they were not used in the training process, which is the research gap addressed in this paper.

3 Methodology

This section presents our proposed PW-MAD approach along with a set of baseline approaches.

3.1 The proposed PW-MAD

Our proposed PW-MAD solution takes advantage of pixel-level supervision, i.e. a label of attack or bona fide for each image pixel, rather than being only supervised by one label for the whole image. This enhances the ability of the algorithm to distribute (spatially) its focus on image areas where the effect of the manipulation is more evident. This is performed with the aim of bringing less focus on non-attack-related artifacts related to the attack building process or the post-processing of the images, and thus enhance the generalizability of the MAD decision. Such a supervision approach has been shown to gain these benefits when dealing with the generalizability of detecting iris and face spoofing attacks [19,15], however, it was never applied to the generalizability sensitive MAD. A simplified illustration of the pixel-wise supervision concept is illustrated in Figure 1.

Our PW-MAD utilizes a densely connected network framework for MAD with binary and deep pixel-wise supervision. This framework is based on the DenseNet [20] architecture. Specifically, we use the DenseNet-121 architecture [20]. The use of this architecture is motivated by the high performances achieved in detecting iris and face spoofing attacks [19,15].

The architecture is modified to be simpler with only two dense blocks and two transition blocks with a fully connected layer with sigmoid activation to produce the binary output. In addition, a convolution layer with a kernel size of 1×1 is added before this fully connected layer, to generate the feature map for pixel-wise supervision. The feature map (size of 14 in our case) generated from this convolution layer is used to supervise the training of the network in a

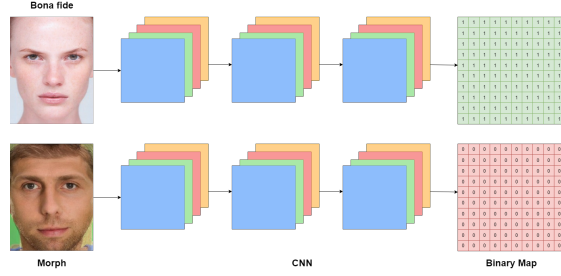


Fig. 1: A simplified illustration of the pixel-wise supervision concept. The target labels (although the same for all pixels per image) are considered for each pixel separately in the training process, rather than a single training target per image.

pixel-wise manner. Finally, the network is trained under pixel-wise and binary supervision. For the loss function, Binary Cross-Entropy (BCE) is used for both pixel-wise and binary supervision. The equation of BCE is:

$$\mathcal{L}_{BCE} = -[y \cdot \log x + (1 - y) \cdot \log(1 - x)]$$

where y presents the ground truth label. x is predicted probability. We use the \mathcal{L}_{BCE}^{PW} to indicate the loss computed based on pixel-wise feature map and \mathcal{L}_{BCE}^B is the loss computed based on binary output. Thus, an overall loss $\mathcal{L}_{overall}$ is formulated as

$$\mathcal{L}_{overall} = \lambda \cdot \mathcal{L}_{BCE}^{PW} + (1 - \lambda) \cdot \mathcal{L}_{BCE}^B,$$

where λ is set to 0.5 in the experiments.

Furthermore, we use the data augmentation technique (horizontal flip, random jitter in brightness, contrast, and saturation) and the hyper-parameters (Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-5}) as motivated in [19] for the training. In addition to data augmentation, we apply class weight and early stopping techniques to avoid overfitting. The final score for each test image is computed by binary output.

3.2 Baselines

LBP: The local binary patterns (LBP) are used extensively in MAD solutions with satisfactory results [38,9]. The face in a frame is first detected, cropped, and normalized into a size of 64×64 pixels. Then, an RGB face is converted into HSV and YCbCr color spaces. Third, the LBP features are extracted from each channel. The multi-channel extraction of LBP features has been shown to enhance the performance of MAD in a number of previous works [28,31]. The obtained six LBP feature vectors are then concatenated into one feature vector to feed into a Softmax classifier, resulting in a decision score.

VGG16: The VGG16 architecture [36] is used extensively in MAD solutions with very competitive results [17,29]. The used network is pretrained on large-scale ImageNet dataset [13] and provided as a pretrained network in [36]. Before

processing the image, it is normalized to 224×224 pixels, then extracts the output of an intermediate layer of VGG16 which is used as a feature. The features are scaled before they are fed to a linear SVM classifier.

Inception_{FT} and **Inception_{TFS}**: This baseline uses the Inception-v3 [39] network architecture as the cornerstone. This architecture has been used successfully for MAD [30] and fake face detection approaches [23]. We report the results of a fine-tuned version of the pre-trained Inception-v3, this will be referred to as Inception_{FT} (the pre-trained network is trained on ImageNet dataset [13] and made available by [39]). The last classification layer of Inception-v3 is modified to fit our two classes case where an input image is either bona fide or attack. Only the weight of this classification layer is fine-tuned, while the weights of other layers are fixed. We also report the results of a trained from scratch Inception-v3 model, named Inception_{TFS}. In the training phase, the binary cross-entropy loss function and Adam optimizer with a learning rate of 10^{-4} and a weight decay of 10^{-5} are used. Moreover, the data augmentation and early stopping techniques used in our PW-MAD method are also applied for training of Inception_{FT} and Inception_{TFS} to avoid overfitting and for a fair comparison.

It must be noted that for PW-MAD and all the baseline solutions used only the training part of the data (and the development split for validation when training a neural network). All the three splits, train, development, and test, are identity-disjoint. All the details of the presented dataset is under Section 4.1 and Table 1

4 Experimental setup

This section presents our newly created morphing attack dataset (with vulnerability analyses) along with the experimental settings and evaluation metrics.

4.1 The dataset

As there is no suitable publicly available morphing dataset, we opted to create a carefully designed morphing dataset that is described in this section. This dataset will be referred to as the digital and re-digitized landmark-based morph dataset (LMA-DRD). The dataset is built on the VGGFace-2 dataset [4], which is composed of 3.31 million images of 9131 identities. This basic dataset was chosen as it has a large number of images per subject, which allows the choice of high-quality samples as will be explained in this section. The images are not scaled and therefore have different resolutions, however as will be clarified, we chose high-resolution images. To cover the frontal image condition in the International Civil Aviation Organisation (ICAO) travel document requirements [21], all non-frontal images are filtered out by detecting the central coordinate of the eyes and the upper coordinate of the nose. The two distances between each of the two eyes and the nose landmarks are calculated, and if the ratio of the difference between these distances to any of them was more than 0.05, the image is neglected. The detected landmarks were used to ensure that all the considered images had an eye-to-eye distance of at least 90 pixels as defined in [21]. Based on these criteria, the total number of images after filtering was 54010 images. This cleaned version

of the data is the one that all the samples in our LMA-DRD dataset originate from.

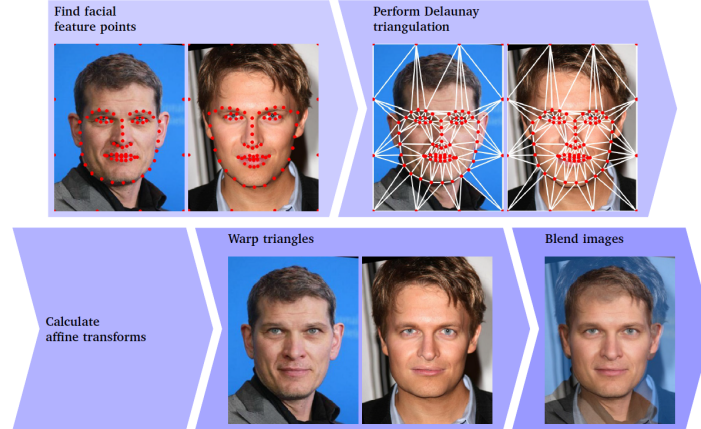


Fig. 2: Face Morphing pipeline followed in this work and described in [28]. The pipeline starts with detecting landmarks in the original (to be morphed) images, and ends up with the blended morphed image.

From the filtered data, and as a starting point, 197 images of 197 identities were manually chosen so that they are split evenly between males and females, frontal faces, with a neutral expression, have no glasses, good illumination quality, and no occlusion. Each of these images was paired with the two most similar faces of two different images of different identities. This is the typically recommended protocol, which makes sense if the goal is to create a confident attack [10]. This pairing depended on the similarity between the key image and the selected paired images. The similarity was measured by the Euclidean distance between the OpenFace representations [2]. The 197 key images were paired twice, resulting in 394 morphing pairs. The paired images (besides the key ones) were not paired with more than one key image. For each of these pairs, a morphed image has been created using the landmark-based approach and parameters presented in [28], the morphing pipeline is illustrated in Figure 2. The 394 morphed images were manually inspected and any image with strong artifacts was regarded, this resulted in a final 276 morphing attack images. From the same identities involved in the attacks, a second bona fide image was chosen (total 591 images), these were manually filtered for quality as described above to comply with ICAO standards. After filtering, the remaining images were 364 images. These 364 images are considered bona fide samples. In total, the created morphing LMA-DRD dataset contains 364 digital Bona fide (D-BF) images and 276 digital morphing attacks (D-M). These images were printed on 11,5cmx9cm glossy photo paper in a professional studio and scanned with 600dpi scanner. They resulted in the same number of re-digitized bona fide (PS-BF) and attacks (PS-M). The resulted LMA-DRD dataset is split into three

identity-disjoint parts, train, development, and test, splits. The splits are done so that they are identity-disjoint, have a similar number of samples, and are equally distributed over males and females to the best possible extend. In our experiments, the training uses only the training data split, the validation during the training (when training a neural network) uses only the development set, and the evaluation is performed only on the test set. The vulnerability analyses in the next paragraph are performed on the the three data splits (train, develop, and test) as the analyses do not include any training. Table 1 present an overview of our LMA-DRD dataset and its splits. Samples of the images included in the dataset and morphing results are shown in Fig. 3.

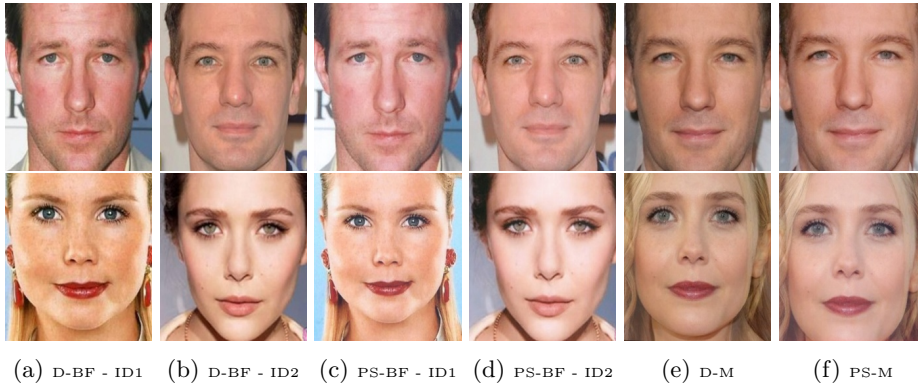


Fig. 3: Samples of our LMA-DRD database with the digital bona fide samples ((a) and (b)), re-digitized bona fide ((d) and (d)), digital morphing attacks (e), and re-digitized morphing attacks (f).

	Attacks		Bona fide	
	Digital (D-M)	Re-digitized (PS-M)	Digital (D-BF)	Re-digitized (PS-BF)
Train	96	96	121	121
Development	92	92	120	120
Test	88	88	123	123
Total	276	276	364	364

Table 1: A detailed view of the presented LMA-DRD database. The numbers indicate the number of images in each data type and data split. Note that the training, development, and testing splits of the data are all identity-disjoint.

The vulnerability (to the morphing attacks) of a pre-trained ResNet-100 ArcFace [14] face recognition model is analysed in this work. ResNet-100 ArcFace [14] pre-trained face recognition model is used as it is one of the most widely used and best performing academic face recognition models with a publicly available

pre-trained network. The vulnerability is measured as the Mated Morph Presentation Match Rate (MMPMR) (as defined in [32]) and is presented in Table 2 for a false match rate (FMR) of 0.1% (as recommended for border check operations by Frontex [18]) and 1.0%, which proves the validity of the considered attacks. An MMPMR of 91.30% at FMR of 1.0% means that 91.30% of the attacks will be matched to both contributing identities if the considered face recognition solution uses the decision threshold at 1.0% FMR. We notice, in table 2, that the vulnerability to re-digitized attacks is slightly less than it is to digital attacks. This might be due to the image artifacts introduced in the re-digitization process. We additionally provide a visual illustration of the face recognition vulnerability to the attacks in Figure 4. The figures plot the similarity score between the attacks (M-D in 4.a and M-PS in 4.b) and the first involved identity (x-axis) vs. the one with the second identity (y-axis). The red lines in these plots represent the threshold value that achieves FMR of 0.1% as recommended for border check operation by Frontex [18]. This helps to put the plotted scores in perspective knowing that any attack represented by a dot in the figure successfully match both identities at this threshold ($\text{FMR} = 0.1\%$) if it is above and to the right of the red lines. The plots in Figure 4, confirm the MMPMR values in Table 2 by showing the high vulnerability of the face recognition system to the presented attacks and the slight drop in this vulnerability after the re-digitization process.

The LMA-DRD data will be made publicly available to researchers upon request. Currently, to the best of our knowledge, this is the first and only face morphing attack dataset with re-digitized images to be made publicly available for academic researchers.

4.2 Experiments and evaluation metrics

Attack	at FMR=0.1%	at FMR=1%
	MMPMR(%)	MMPMR(%)
D-M	80.07	91.30
PS-M	77.17	88.41

Table 2: The vulnerability to the LMA-DRD dataset attacks, both the digital (D-M) and the re-digitized (PS-M) represented by the MMPMR(%) at two different decision thresholds ($\text{FMR}=0.1\%$ and 1.0%) of the investigated ResNet-100 Arcface pre-trained model. Note the slight decrease in vulnerability to the re-digitized attacks when compared to the digital ones.

Our experiments aim at evaluating the generalizability of our proposed PW-MAD and the other baseline MADs. As baseline experiments, we evaluate the different MADs on the same type of data (digital or re-digitized). This results in two baseline experimental settings, one uses the digital data for training and testing (Train-D Test-D) and one uses the re-digitized data for training and testing (Train-PS Test-PS). Two additional experimental setups measure the generalizability on data of an unknown type. One uses the digital data for training

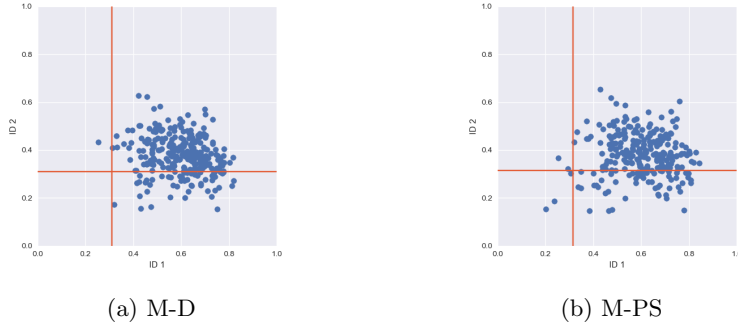


Fig. 4: The vulnerability of face recognition to the attacks in the LMA-DRD dataset attacks is represented by the similarity of the attack to the two identities used to create the attack (ID1 on the X-axis and ID2 on the y-axis). The red lines represent the similarity threshold for the FMR of 0.1%, which indicates that all the attacks on the top and to the right of the red lines do match the targeted identities at this FMR setting. Plot (a) represents the digital attacks, and (b) represents the re-digitized attacks.

and re-digitized data for testing (Train-D Test-PS) and one uses the re-digitized data for training and digital data for testing (Train-PS Test-D). It must be noted that the "Train-D Test-PS" reflects the most application-relevant use-case and thus the most commonly reported case on MAD generalization in the literature [27,29,17]. In our experiments, the training uses only the training data split, the validation during the training (when training a neural network) uses only the development set, and the evaluation is performed only on the test set. The three sets are identity-disjoint to prevent biases in the evaluation.

The MAD performance is presented by the Attack Presentation Classification Error Rate (APCER), i.e. the proportion of attack images incorrectly classified as bona fide samples, and the Bona fide Presentation Classification Error Rate (BPCER), i.e. the proportion of bona fide images incorrectly classified as attack samples, as defined in the ISO/IEC 30107-3 [22]. To cover different operation points, and to present the comparative results, we report the BPCER at three different fixed APCER values (0.1%, 1.0%, and 10.0%). To provide a visual evaluation on a wider operation range, we plot receiver operating characteristic (ROC) curves by plotting the APCER on the x-axis and 1-BPCER on the y-axis at different operational points. It must be noted again that the MAD evaluation was performed only on the identity-disjoint test data as described in Section 4.1.

5 Results and discussion

Table 3 lists the BPCER rates achieved at different APCER thresholds for the PW-MAD and the baseline MADs. On the intra-data type settings (Train D Test D) the PW-MAD solution outperforms all baselines at the lowest APCER operation point (0.1%) by scoring a BPCER of 17.74% in comparison to 34.67% for the next best MAD. For higher APCER values, the PW-MAD scores the

second-best BPCER. Also for the intra-data type settings (Train PS Test PS), the proposed PW-MAD scores the lowest BPCER (best) on all operational points (APCER thresholds). This is supported by the ROC curves in Figures 5.a and 5.b.

In the more challenging inter-data type settings, the proposed PW-MAD outperformed all the baselines at all the APCER thresholds. For the "Train-D Test-PS" setting, the BPCER (at APCER of 1.0%) scored by our PW-MAD is 32.52% in comparison to 49.59% for the next best MAD. This constitutes a 34.4% drop in the BPCER value. For the "Train-PS Test-D" setting, the BPCER (at APCER of 1.0%) scored by our PW-MAD is 19.35% in comparison to 51.61% for the next best MAD. This constitutes a 62.5% drop in the BPCER value. These inter-data type evaluation results demonstrate the superior generalizability of our proposed PW-MAD in comparison to the baselines. These inter-data type conclusions are supported by the ROC curves in Figures 5.d and 5.e, where the better maintenance of the performance (in comparison to the baselines) is apparent when comparing these curves to the ones in Figures 5.a and 5.b.

Approach	Train	Test	BPCER (%) @ APCER =			Approach	Train	Test	BPCER (%) @ APCER =		
			0.1%	1.0%	10.0%				0.1%	1%	10%
LBP	D	D	51.61	51.61	25.80	LBP	PS	PS	98.37	95.93	57.72
VGG16	D	D	67.74	50.00	32.25	VGG16	PS	PS	63.41	54.47	39.83
Incep. <i>FT</i>	D	D	98.38	97.58	81.45	Incep. <i>FT</i>	PS	PS	97.56	97.56	91.05
Incep. <i>TFS</i>	D	D	34.67	12.09	4.83	Incep. <i>TFS</i>	PS	PS	71.54	48.78	13.00
PW-MAD	D	D	17.74	16.12	6.45	PW-MAD	PS	PS	31.70	31.70	8.13
LBP	D	PS	96.74	80.48	51.21	LBP	PS	D	75.80	75.00	46.77
VGG16	D	PS	80.48	74.79	40.65	VGG16	PS	D	70.96	51.61	33.06
Incep. <i>FT</i>	D	PS	95.93	94.30	62.60	Incep. <i>FT</i>	PS	D	95.16	86.29	79.03
Incep. <i>TFS</i>	D	PS	69.91	49.59	28.45	Incep. <i>TFS</i>	PS	D	81.45	62.09	25.80
PW-MAD	D	PS	40.65	32.52	12.19	PW-MAD	PS	D	66.12	19.35	6.45

Table 3: The BPCER at different APCER values (the lower BPCER, the better the MAD performance) achieved by our PW-MAD and the different considered baselines. One can note the better performance of the proposed PW-MAD on most experimental settings, especially when considering the realistic cross-attack scenario on the bottom two tables. Train and Test indicate the data type used for training and testing (digital (D) or re-digitized (PS)). The lowest BPCER for each train/test setup is in bold for each APCER threshold.

To put the generalizability of the proposed approach in perspective, we compare our PW-MAD with the previously published approaches that targeted the detection of re-digitized morphing attacks [29,17,28]. As these works reported their results on private datasets, we are not able to build a direct performance comparison. We rather present the reported results when the MAD is trained and tested on digital morphs, along with the performance of the MAD trained on digital morphs and tested on re-digitized morphs, see Table 4 where the performances are reported in BPCER at APCER of 10% as it is the common reported measure between the relevant previous works. We also list the BPCER error increase (in percentage points) when moving from testing on the known digital

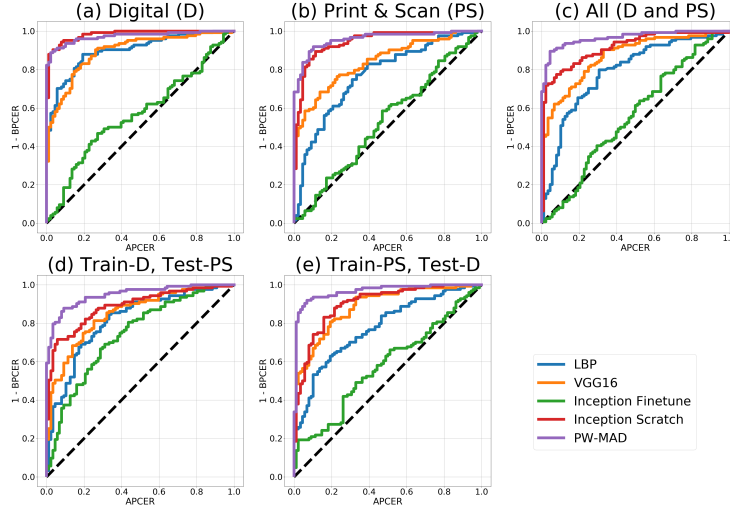


Fig. 5: ROC curves achieved by our PW-MAD solution along with the other baselines. The considered experimental settings are Train-D Test-D in (a), Train-PS Test-PS in (b), Train-D and PS, Test-D and PS in c(), Train-D Test-P in (d), and Train-PS Test-D in (e). Note the superior performance of the PW-MAD, especially in the inter-data type settings in (d) and (e).

morphs to the unknown re-digitized. This performance drop might not be an optimal measure of the performance, as it neglects the absolute performance, but it rather gives a clear indication of the generalization. It is noted in Table 4 that our proposed PW-MAD results in the lowest performance drop between previously reported results, indicating the relatively high generalizability of its decisions. When it comes to training on re-digitized attacks and testing on digital attacks, our PW-MAD actually gains performance, BPCER at 10% APCER moves from 8.13% to 6.45% as in Table 3. This training/testing protocol was only reported in previous literature in [29], where their transfer learning approach reported in the best-case scenario, a BPCER at 10% APCER of 16.43% on known re-digitized attacks and dropping to 30.13% when testing on the unknown digital attacks. This again points out the relative generalizability of our proposed PW-MAD approach.

6 Conclusion

This work targeted the enhancement of the generalizability of MAD performance. This is achieved by proposing the PW-MAD solution that leverages the adaption of pixel-wise supervision into the training process to produce a stable performance, even when facing unknown variations like re-digitized images. We presented a new dataset that included digital and re-digitized samples, allowing the development and evaluation of the proposed PW-MAD. The PW-MAD

Approach	BPCER (%) @ APCER = 10%		BPCER increase in percentage points
	Train-D, Test-D	Train-D, Test-PS	
BSIF-SVM (P1) [27,29]	38.25	48.63	10.38
BSIF-SVM (P2) [27,29]	38.25	57.53	19.28
Transferable D-CNN (P1) [29]	7.53	24.65	17.12
Transferable D-CNN (P2) [29]	7.53	17.8	10.27
Fine-tune AlexNet [17]	0.8	50.8	50
Fine-tune VGG19 [17]	0.8	32.7	31.9
Fine-tune VGG-Face16 [17]	0.8	13.8	13
Fine-tune VGG-Face2 [17]	0.0	20	20
PW-MAD (ours)	6.45	12.19	5.74

Table 4: A comparison on the results presented in the state-of-the-art works reporting on experimental settings where the MAD is trained and tested on digital attacks (Train-D, Test-D) and when trained on digital attacks and tested on re-digitized attacks (Train-D, Test-PS). The BPCER values are not directly comparable, as each of the works considered a different (private) database. The increase in BPCER percentage points represents the generalizability of the MAD to unknown variations in the attack and it shows that our proposed PW-MAD achieves the lowest drop in the performance, and thus relatively high generalizability. The lowest increase (percentage points) in BPCER error between the two experimental setups is in bold. P1 and P2 indicate using different printers in the respective papers.

proved to provide a superior MAD generalizability over a set of widely used baselines and previously reported results in state-of-the-art.

Acknowledgment: This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

1. Aghdaie, P., Chaudhary, B., Soleymani, S., Dawson, J.M., Nasrabadi, N.M.: Detection of morphed face images using discriminative wavelet sub-bands. In: IWBF. pp. 1–6. IEEE (2021)
2. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science (2016)
3. Bolle, R., Pankanti, S.: Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society. Kluwer Academic Publishers, Norwell, MA, USA (1998)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG. pp. 67–74. IEEE Computer Society (2018)

5. Damer, N., Boller, V., Wainakh, Y., Boutros, F., Terhörst, P., Braun, A., Kuijper, A.: Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts. In: GCPR. Lecture Notes in Computer Science, vol. 11269, pp. 518–534. Springer (2018)
6. Damer, N., Boutros, F., Saladie, A.M., Kirchbuchner, F., Kuijper, A.: Realistic dreams: Cascaded enhancement of gan-generated images with an example in face morphing attacks. In: BTAS. pp. 1–10. IEEE (2019)
7. Damer, N., Dimitrov, K.: Practical view on face presentation attack detection. In: BMVC. BMVA Press (2016)
8. Damer, N., Grebe, J.H., Zienert, S., Kirchbuchner, F., Kuijper, A.: On the generalization of detecting face morphing attacks as anomalies: Novelty vs. outlier detection. In: BTAS. pp. 1–5. IEEE (2019)
9. Damer, N., Saladie, A.M., Braun, A., Kuijper, A.: MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network. In: BTAS. pp. 1–10. IEEE (2018)
10. Damer, N., Saladie, A.M., Zienert, S., Wainakh, Y., Terhörst, P., Kirchbuchner, F., Kuijper, A.: To detect or not to detect: The right faces to morph. In: ICB. pp. 1–8. IEEE (2019)
11. Damer, N., Zienert, S., Wainakh, Y., Saladie, A.M., Kirchbuchner, F., Kuijper, A.: A multi-detector solution towards an accurate and generalized detection of face morphing attacks. In: FUSION. pp. 1–8. IEEE (2019)
12. Debiasi, L., Damer, N., Saladie, A.M., Rathgeb, C., Scherhag, U., Busch, C., Kirchbuchner, F., Uhl, A.: On the detection of gan-based face morphs using established morph detectors. In: ICIAP (2). Lecture Notes in Computer Science, vol. 11752, pp. 345–356. Springer (2019)
13. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE Computer Society (2009)
14. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR. pp. 4690–4699. Computer Vision Foundation / IEEE (2019)
15. Fang, M., Damer, N., Boutros, F., Kirchbuchner, F., Kuijper, A.: Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In: IJCB. pp. 1–8. IEEE (2021)
16. Ferrara, M., Franco, A., Maltoni, D.: The magic passport. In: IJCB. pp. 1–7. IEEE (2014)
17. Ferrara, M., Franco, A., Maltoni, D.: Face morphing detection in the presence of printing/scanning and heterogeneous image sources. IET Biometrics **10**(3), 290–303 (2021)
18. Frontex: Best practice technical guidelines for automated border control (abc) systems (2015)
19. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019. pp. 1–8. IEEE (2019)
20. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 2261–2269. IEEE Computer Society (2017)
21. International Civil Aviation Organization, ICAO: Machine readable passports – part 9 – deployment of biometric identification and electronic storage of data in eMRTDs. Civil Aviation Organization (ICAO) (2015)

22. International Organization for Standardization: ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting (2017)
23. Khodabakhsh, A., Raghavendra, R., Raja, K.B., Wasnik, P.S., Busch, C.: Fake face detection methods: Can they be generalized? In: BIOSIG. LNI, vol. P-282, pp. 1–6. GI / IEEE (2018)
24. Makrushin, A., Neubert, T., Dittmann, J.: Automatic generation and detection of visually faultless facial morphs. In: VISIGRAPP (6: VISAPP). pp. 39–50. SciTePress (2017)
25. Markets and Markets: Facial Recognition Market by Component (Software Tools and Services), Technology, Use Case (Emotion Recognition, Attendance Tracking and Monitoring, Access Control, Law Enforcement), End-User, and Region - Global Forecast to 2022. Report (November 2017)
26. Massoli, F.V., Carrara, F., Amato, G., Falchi, F.: Detection of face recognition adversarial attacks. *Comput. Vis. Image Underst.* **202**, 103103 (2021)
27. Raghavendra, R., Raja, K.B., Busch, C.: Detecting morphed face images. In: BTAS. pp. 1–7. IEEE (2016)
28. Raghavendra, R., Raja, K.B., Venkatesh, S., Busch, C.: Face morphing versus face averaging: Vulnerability and detection. In: IJCB. pp. 555–563. IEEE (2017)
29. Raghavendra, R., Raja, K.B., Venkatesh, S., Busch, C.: Transferable deep-cnn features for detecting digital and print-scanned morphed face images. In: CVPR Workshops. pp. 1822–1830. IEEE Computer Society (2017)
30. Ramachandra, R., Venkatesh, S., Raja, K.B., Busch, C.: Detecting face morphing attacks with collaborative representation of steerable features. In: CVIP (1). *Advances in Intelligent Systems and Computing*, vol. 1022, pp. 255–265. Springer (2018)
31. Ramachandra, R., Venkatesh, S., Raja, K.B., Busch, C.: Towards making morphing attack detection robust using hybrid scale-space colour texture features. In: ISBA. pp. 1–8. IEEE (2019)
32. Scherhag, U., Nautsch, A., Rathgeb, C., Gomez-Barrero, M., Veldhuis, R.N.J., Spreeuwens, L.J., Schils, M., Maltoni, D., Grother, P., Marcel, S., Breithaupt, R., Raghavendra, R., Busch, C.: Biometric systems under morphing attacks: Assessment of morphing techniques and vulnerability reporting. In: BIOSIG. LNI, vol. P-270, pp. 149–159. GI / IEEE (2017)
33. Scherhag, U., Raghavendra, R., Raja, K.B., Gomez-Barrero, M., Rathgeb, C., Busch, C.: On the vulnerability of face recognition systems towards morphed face attacks. In: IWBF. pp. 1–6. IEEE (2017)
34. Scherhag, U., Rathgeb, C., Busch, C.: Performance variation of morphed face image detection algorithms across different datasets. In: IWBF. pp. 1–6. IEEE (2018)
35. Scherhag, U., Rathgeb, C., Merkle, J., Busch, C.: Deep face representations for differential morphing attack detection. *IEEE Trans. Inf. Forensics Secur.* **15**, 3625–3639 (2020)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
37. Soleymani, S., Chaudhary, B., Dabouei, A., Dawson, J.M., Nasrabadi, N.M.: Differential morphed face detection using deep siamese networks. In: ICPR Workshops (6). *Lecture Notes in Computer Science*, vol. 12666, pp. 560–572. Springer (2020)
38. Spreeuwens, L.J., Schils, M., Veldhuis, R.N.J.: Towards robust evaluation of face morphing detection. In: EUSIPCO. pp. 1027–1031. IEEE (2018)

39. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2818–2826. IEEE Computer Society (2016)
40. Venkatesh, S., Ramachandra, R., Raja, K., Busch, C.: Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society* **2**(3), 128–145 (2021). <https://doi.org/10.1109/TTS.2021.3066254>
41. Zhang, H., Venkatesh, S., Ramachandra, R., Raja, K.B., Damer, N., Busch, C.: MIPGAN - generating strong and high quality morphing attacks using identity prior driven GAN. *IEEE Trans. Biom. Behav. Identity Sci.* **3**(3), 365–383 (2021)