Image and Vision Computing 126 (2022) 104535

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Towards generalized morphing attack detection by learning residuals☆

Kiran Raja *, Gourav Gupta, Sushma Venkatesh, Raghavendra Ramachandra, Christoph Busch

Norwegian University of Science and Technology, Norway

ARTICLE INFO

Article history:
Received 29 December 2021
Received in revised form 9 August 2022
Accepted 11 August 2022
Available online 28 August 2022

Keywords:
Morphing attacks
Morphing attack detection
Face recognition
Vulnerability of biometric systems

ABSTRACT

Face recognition systems (FRS) are vulnerable to different kinds of attacks. Morphing attack combines multiple face images to obtain a single face image that can verify equally against all contributing subjects. Various Morphing Attack Detection (MAD) algorithms have been proposed in recent years albeit limited generalizability. We present a new approach for MAD in this work with better generalization than state-of-the-art (SOTA) algorithms. We propose an end-to-end multi-stage encoder-decoder network for learning the residuals of morphing process to detect attacks. Leveraging the residuals, we learn an efficient classifier using cross-entropy loss and asymmetric loss. The use of asymmetric loss in our approach is motivated by imbalanced distribution of morphs and bona fides. An extensive set of experiments are conducted on five different datasets consisting of two landmark based and three Generative Adversarial Network (GAN) based morphs in various settings such as digital, print-scan and print-scan-compression. We first demonstrate a near-ideal performance of the proposed MAD with Detection Equal Error Rate (D-EER) of 0% in the best case and 2.58% in the worst case in the digital domain in closed-set protocol, i.e., known attacks. Further, we demonstrate the applicability of the proposed approach on 60 different combinations where the testing set contains unknown morphing attacks in open-set protocol to illustrate the generalization ability of our proposed approach. Through training the proposed approach on landmark-based morph generation data alone, we obtain an EER of 3.59% in the best case and 12.89% in the worst case for morphed images in the digital domain, reducing the error rates from 45.67% and 30.23% respectively, in open-set protocol. We further present an extensive analysis of the proposed approach through Class Activation Maps (CAM) to explain the decisions using by making use of three complementary CAM analysis.

1. Introduction

With recent progress in image manipulation techniques, the attacks and the attack potential is increasing making the FRSs vulnerable. Recently studied morphing attacks combine face images of multiple subjects to obtain a single image [1–7]. A resulting image, commonly referred to as a morphed image, compromises FRSs by providing a high comparison score to all contributing face images represented in morphed face image. For instance, a malicious actor can obtain a valid ID and carry out illegal activities if the morphed image is uploaded to obtain ID cards leading to security lapses. It is therefore critical to have attack detection, especially in unsupervised access control settings.

A number of recent works have proposed Morphing Attack Detection (MAD) algorithms to mitigate the attacks. MAD algorithms either determine a given image as a bona fide or morphed image using different approaches making use of classical machine learning or end-to-end learned deep networks [1–6]. MAD approaches further are designed to consider scenarios where a single image has to be decided as a bona fide or morphed image and are categorized under Single-Image MAD (S-MAD) [8]. Similarly, MAD algorithms are also proposed when a reference image is available to detect morphing attacks (i.e., Differential-MAD (D-MAD)) [8].

The recent advances enable the creation of morphed attacks with minimal efforts using GANs [9–12] while challenging to detect if such samples are not seen during training. Detecting the attacks is further challenging if the images after morphing are post-processed, printed and scanned (re-digitalized) [3,6,8] or printed-scanned and compressed [8,9,13]. Further, as all possible morph generation cannot be known at training phase, the problem can be posed as an open set detection task. The set of all related works are listed in the section below for the convenience of the reader where we note limited generalizability of existing works addressing unknown attacks.

2. Related works on S-MAD

The initial set of MAD algorithms focused on detecting the attacks in digital domain and illustrated a very high attack detection rate [2,11,14,

Abbreviations: MAD, Morphing Attack Detection; S-MAD, Single Image Morphing Attack Detection; D-MAD, Differential Morphing Attack Detection.

☆ This work was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 883356.

* Corresponding author.
E-mail address: kiran.raja@ntnu.no (K. Raja).

https://doi.org/10.1016/j.imavis.2022.104535

15–21]. These algorithms were based on hand-crafted features and focused on localizing and detecting the artefacts at the image level reflected at digital pixel-level information [2,14,15–19]. In all of these works, a classifier was trained on the texture-level and observing pixel-level features to detect morph attacks. Noting the poor visual appearance of the morphed images from initial morphing approaches using landmarks, subsequent works proposed approaches which can better control the quality of morphed images by carefully choosing the contributing pairs in the same ethnic group, age, and gender [3,9,20,21]. Further, a number of post-processing steps have been used, which typically involve retouching the images to make the hair silhouette visually realistic, eliminating the artefacts due to incorrect registration of the iris region and making the skin colour consistent. Generative Adversarial Networks (GAN) have also been explored to create realistic images with ease [9,11]. With the improved visual quality of morphed images challenging human perception, one can foresee the need for better MAD algorithms. A number of new MAD approaches have therefore been proposed using deep learning [3,8,22–24]. All existing works in this direction are typically based on pre-trained networks and transfer learning. The first work in this direction was based on using pre-trained networks such as AlexNet and VGG18, whose features are fused and classified to detect a morphing attack [3]. Following this, several deep CNN pre-trained networks such as AlexNet, VGG19, VGG-Face16, GoogleNet, ResNet18, ResNet150, ResNet50, VGG-Face2 and Open face [21,25,26,27–32] have been explored.

While deep networks have improved MAD performance as compared to hand-crafted MAD methods on both digital and print-scan data, the generalization capability of these approaches is limited across different print-scan datasets [6,9,22]. Venkatesh et al. [21] proposed using multi-scale Context Aggregation Networks (CAN) to detect the morphing traces from images and used it efficiently to detect the morphing attacks. Scherhag et al. [22] proposed using the embeddings from deeply learnt FRS to detect the morphing attacks and demonstrated the applicability on two independent evaluations conducted by a European project [8,13] and the NIST FRVT MORPH evaluation [33].

In addition to this, wavelet-based approaches for detecting morphs have been proposed using attention aware mechanisms [23], group sparsity [34] and mutual information [24]. Approaches have explored Siamese networks [35] for detecting morphs and feature space from morphs [36]. Most recently, adversarial learning has been explored to make the MAD algorithms better by using adversarial examples [37,38]. Pixel level information has been explored, specifically to detect morphs in the digital domain [39].

In an alternative direction, hybrid approaches have been proposed combining more than one feature extractor or classifier to detect face morphing attacks in a better manner. Approaches have been earlier proposed to combine features and detection or decision scores for better MAD accuracy [20,26,28,31,40,41]. As these approaches combine more than one feature extraction and classifier, the MAD performance is generally superior to single-image MAD techniques. However, while they provide better MAD performance when tested on known data (i.e., closed-set protocol), they cannot generalize well to detect attacks from different types of morphs and morphs after print-scan processes. Further, the robustness of the MAD algorithms in multiple works has been tested on both internal datasets with an exception of few that have been tested on sequestered datasets in NIST FRVT MORPH evaluation [33], and SOTAMD evaluation [13,42].

Retrospection of the MAD algorithms reported so far in the literature leads to two critical observations in lines with observations of other recent works [6,8,9]:

- MAD algorithms tend to perform well when trained and tested with the same type of morphing data, for instance, digital morphed images against digital bona fide images created using a specific kind of morphing (i.e., closed-set protocol). However, the algorithms suffer from performance degradation when presented with morphed images from unseen algorithms (i.e., open-set protocol).
- The performance degradation is further pronounced when the morphed images of a specific generation process are tested against a model trained on different morphed images, especially in printed-and-scanned and printed-scanned-compressed settings.

Both of these factors in MAD in an open-set scenario (images from different morph types in testing than in the training phase) hinder the deployment of MAD algorithms owing to high error rates for Attack Presentation Classification Error Rate (APCER) at a fixed Bona fide Classification Error Rate (BPCER). The degradation in performance can be seen in a recent work reported by Zhang et al. [9] where the same morph type generation for testing and training resulted in high detection accuracy close to 0% Equal Error Rate (EER) while it degraded in cross-generation type testing. The performance of EER degradation is shown in Fig. 1 where one can see that the unknown morphing generation in the testing set results in a higher Equal Error Rate (EER) in SOTA methods. At the same time, the proposed approach reduces the EER and BPCER, as noted in Fig. 1.

2.1. Our contributions

Considering the challenges in cross-data MAD, we propose a new approach for detecting morphing attacks by incorporating three basic but necessary ideas to make MAD algorithms better.

- We note that the morphed image and the bona fide image, despite looking very similar in RGB colour space, present complementary information in other colour spaces as illustrated in Fig. 2. We thus assert that using such complementary colour information from different colour spaces can lead to better MAD algorithms. Our motivation also stems from the capture devices, printers and scanners operating on various colour gamuts, adjusting the colour spaces accordingly. While previous work [20,40] proposed MAD algorithm using the independent colour spaces to extract the features and then learn classifiers independently, we assert that learning features from different colour spaces together can lead to increased accuracy in detecting morphs. RGB can serve as a colour representation for a wide variety of capture devices, specifically for bona fide images and morphed images that are printed and scanned using a variety of printers, HSV colour space is device-independent. HSV colour space encapsulates information about a colour in a manner that is more familiar to humans and help in detecting post-processed morphed images where familiar colours are used to eliminate the artefacts such as ghosting artefacts. Lab colour space is further designed to approximate human vision and consider perceptual uniformity. The L component closely matches the human perception of lightness. It can thus be used to make accurate colour balance corrections by modifying output curves in the a and b components or to adjust the lightness contrast using the L component. Multiple colours within Lab space cannot be reproduced in the physical world. Thus if an image editing software is employed, for instance, in morph post-processing, the colours would be the closest in-gamut approximation, changing lightness and colourfulness. We assert that such clues can be used to benefit from morphing attack detection. YCbCr colour space is further based on the RGB colour model and is used for storage and data transmission due to its efficiency in compressing colour data in images and video. Asserting that morphed images can be transmitted through a digital transmission device or when the images are compressed to correspond to passport size, using only RGB representation may lead to not using important clues in MAD. Thus, we propose using four different colour spaces such as RGB, HSV, Lab and YCbCr for creating the MAD algorithm based on the complementarity as illustrated in Fig. 2.
- Secondly, many morph generation algorithms manipulate the image information, and such a process tends to leave traces of the morphing

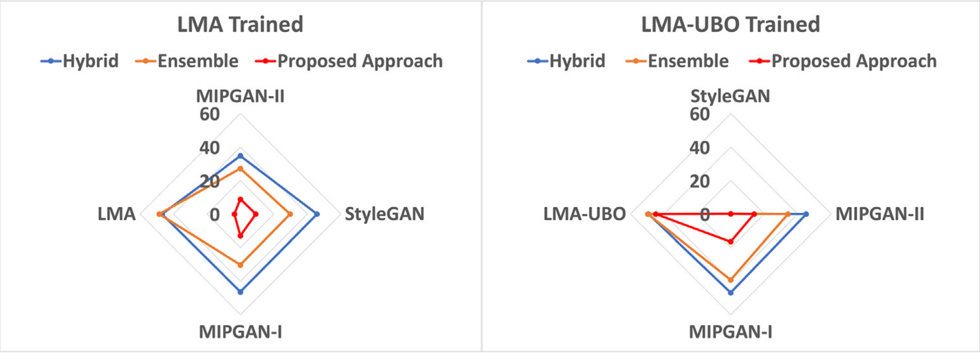


Fig. 1. Illustration of high EER for cross-dataset MAD under training with LMA and LMA-UBO tested with data from unknown morphing generation in digital domain. Higher EER can be noted for SOTA methods such as Ensemble Features [20] and Hybrid Features [41] while proposed approach reduces the EER significantly.

process in the resulting image. While colour spaces can reveal them in complementary spaces, we also assert that using residuals jointly can be advantageous. The problem, however, remains in reliably obtaining the residuals. Therefore, we propose employing an encoder-decoder network to recover the images that approximate the input images, either bona fide or morphs. We hypothesize that the artefacts commonly present in morphs can easily be captured through an encoder-decoder network different from bona fide images. Therefore, we propose using such reconstructed images from the encoder-decoder network as auxiliary information to pronounce these artefacts before learning a classifier as described in the upcoming section.

- We conduct an extensive evaluation on a large scale in–house database consisting of ICAO compliant face images and their corresponding morph images from 5 different types of morph generation algorithms [9]. The evaluation considers morphs generated from two different landmark based approaches (LMA [43], and LMA-UBO [25]) and three different GAN based morphing approaches (StyleGAN [10], MIPGAN-I and MIPGAN-II [9]). Further, bona fide and morphed images from all five datasets are analyzed in digital, print-scan and printed-scan-compression domains to study the generalizability of the proposed approach. In addition, the approach is benchmarked against two SOTA MAD algorithms evaluated in NIST FRVT MORPH evaluation [33].

In the rest of this article, we present the proposed approach in Section 3 with a detailed discussion. We then present the details of experiments in Section 4 with a brief summary of database in Section 4.1. We present the baseline results in Section 4.5 along with a set of results on ablation study in Section 4.6. Further, the results on generalization is presented in Section 4.7 along with a detailed analysis on explainability in Section 5 and conclusion in Section 6.

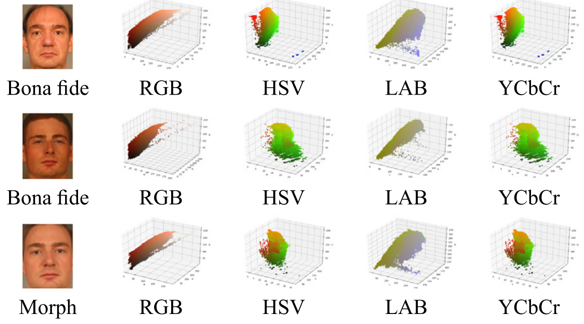


Fig. 2. Illustration of complementarity of color spaces for bona fide and morph images.

3. Proposed approach

The proposed approach is based on our assertion that the bona fide samples have a few properties different from the morphed samples, as argued before. A morphed image can be seen as a noisy version of the bona fide image with multiple residues due to the morphing. We note that post-processing and printing-scanning remains a common factor across bona fide and morphs, while traces of the morphing process are absent in bona fide image.

Suppose I^b_{S1} and I^b_{S2} are bona fide images of two subjects $S1$ and $S2$. In that case, the morphed image I^m can be represented as $I^m \mapsto \phi\{I^b_{S1}, I^b_{S2}\}$ where $\phi\{I^b_{S1}, I^b_{S2}\}$ without loss of generality can be approximated as a function which involves the process of morphing, post-processing and printing-scanning. The function $\phi\{.,.\}$ varies based on the generation mechanisms involved, for instance Land-Mark Based Morphing [43, 25], StyleGAN [10] or MIPGAN [9]. While the latter two involve manipulating the latent space to create morphs, the former involves manipulating the pixel level information in images. Thus, we propose to exploit the traces resulting from the morphing process to devise MAD, i.e., to invert the $\phi\{.,.\}$. Our proposed approach consists of two steps, (i) identifying the traces (which we refer to as residuals for the sake of consistency in the rest of this article) and (ii) learning a classifier to detect morphs against bona fide using the residuals. Diverse morph generation processes result in images that exhibit different image characteristics due to the process of morphing, post-processing, retouching. Considering a compact sphere formed by bona fide images, we can note that morphed images drift farther away from the center of the compact sphere. We assert to create a generalizable MAD approach utilizing this and learning the residuals of the morphed images. We, therefore, detail the process of learning residuals in the first stage, as described in the next section.

3.1. Morph residual learning

In an ideal setting, an encoder-decoder architecture should faithfully reproduce the original image, i.e., given an input x, the encoder-decoder network produces an image $\hat{x} \approx x$. The residual $R = \hat{x} - x$ in such a case is expected to be very low. However, given a number of steps of involved in morphing and post-processing, taking complementary information to compute residuals is an advantage. Considering the morphing process as modelled by Eq. (1):

$$I^m \mapsto \phi\{I^b_{S1}, I^b_{S2}\} \tag{1}$$

the residuals can be approximated by Eq. (2):

$$R \approx \phi\{I^b_{S1}, I^b_{S2}\} - I^m \tag{2}$$

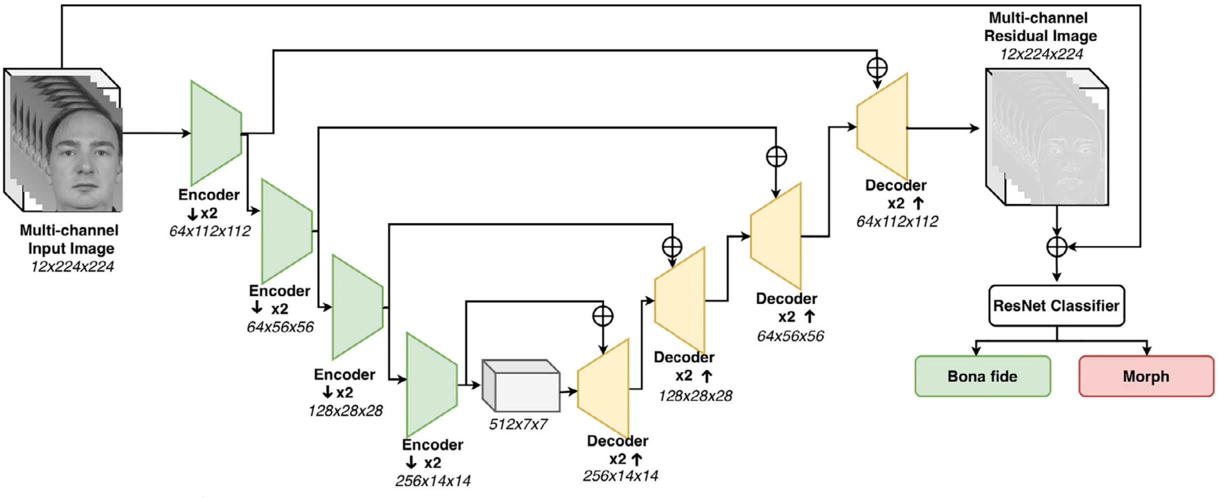


Fig. 3. The proposed network architecture for MAD using U-Net architecture for multi-channel input images with ResNet18 backbone.

Making use of complementary information available in different colour spaces (i.e., different channels) and revise the Eq. (2) as given by Eq. (3).

$$R_c \approx \sum_{i=1}^{c} \phi\{I^b_{S1}, I^b_{S2}\} - I^m \tag{3}$$

where c indicates different channels from different color spaces ($c = 12$ from RGB, HSV, Lab, YCbCr color spaces). We, therefore, learn the residuals from multiple color channels as they differ for bona fide and morphs.

As the first part of our proposed approach, we learn such residuals using the encoder-decoder network using all the 12 channels simultaneously through a multi-channel ResNet [44]. We choose ResNet18 as a backbone network considering its design of skip connection to mitigate the so-called degradation problem [44]. While a simple encoder-decoder can be used, we resort to using a U-Net like architecture with four encoders and four decoders which we refer to as Encoder Residual Units (ERU) encoding the image and Decoder Residual Units (DRU), which reconstruct the images. In each of the DRUs, the feature map from the previous layer is upsampled using nearest-neighbour interpolation. Unlike the classical U-Net architecture, our approach based on multi-channel inputs learns the residuals in different color spaces simultaneously. The reconstructed image (I_c^o) can then be used to obtain the residual difference (R_c) between the input image (I_c^i) and the reconstructed output image (I_c^o) of original input image.

$$R_c \approx I_c^o - I_c^i \tag{4}$$

for all channels c and the input image I^i and output image I^o. As we are interested in learning residual differences for morphs and bona fide images, we obtain the difference between the original image and reconstructed images for both the cases. However, the input image and reconstructed image exhibit different statistical characteristics that can lead to unstable parameters in network and result in vanishing/exploding gradients problems. Thus, we propose to zero-center and normalize the residual difference over the image by employing the mean $\mu(R_c)$ and variance $\sigma^2(R_c)$ for an image to obtain the actual residual difference D_c from Eq. (4) as given in Eq. (5).

$$D_c = \frac{R_c - \mu(R_c)}{\sqrt{\sigma^2(R_c)}} \tag{5}$$

With the ResNet18 as the backbone in our proposed approach, each ERU consists of a convolutional layer followed by batch normalization, rectified linear activation function (ReLU) and max-pool layer. Each convolution layer in ERU consists of a kernel of size 3x3, with a stride of 1 and a padding of 1. Each DRU consists of a block with a convolutional layer of kernel 3 × 3, a stride of 1 and a padding of 1 followed by batch normalization, rectified linear activation function (ReLU) and max-pool layer and another block with kernel 3 × 3, a stride of 1 and a padding of 1 followed by batch normalization, downsampling layer. The output of the first two bocks is then fed to a convolution layer with a kernel of 2 × 2 followed by normalization and ReLU followed by a convolution layer of 1 × 1 and ReLU. Each DRU is designed to upscale the output by an interpolating factor of 2 before passing on to the next DRU, as illustrated in Fig. 3. An off the shelf classifier architecture is used with adaptive average pooling and a linear activation function in a fully connected layer with a drop out of 0.4. We have a total of 4 ERU and 5 DRU in our proposed approach as illustrated in Fig. 3[1].

3.2. Learning to classify morphs

Formulating a morphed sample as a noisy version of a bona fide image, we can assert the bona fide samples belong to a closed-set space while the morphed samples are outliers from this closed-set. Based on such a formulation, we can consider the distribution of the bona fide samples to lie in the center of a compact sphere in the learned feature representation space while the morph samples drift from the center. Considering the input space ($\mathcal{X} \subseteq \mathbb{R}^d$) and output space ($\mathcal{Z} \subseteq \mathbb{R}^p$) and $\phi(\cdot;\mathcal{W}) : \mathcal{X} \rightarrow \mathcal{Z}$ as the neural network with L hidden layers, the corresponding set of weights can be represented as $\mathcal{W} = \{\mathcal{W}^1, \ldots, \mathcal{W}^L\}$. Given N_b bona fide samples $(x_1, \ldots, x_{N_b} \subseteq \mathcal{X}), N_m$ morphed samples $(y_1, \ldots, y_{N_m} \subseteq \mathcal{X})$, let c be the center of the bona fide samples in the output space $\mathcal{Z}$, the objective is:

$$\min_{\mathcal{W}} \frac{1}{N_b} \sum_{i=1}^{N_b} \|\phi(x_i;\mathcal{W}) - c\|^2, \tag{6}$$

$$\max_{\mathcal{W}} \frac{1}{N_m} \sum_{i=1}^{N_m} \|\phi(y_i;\mathcal{W}) - c\|^2. \tag{7}$$

Thus, the distance from $\phi(q;\mathcal{W})$ to the center of the bona fide hypersphere for any image q can be represented as:

$$s(x) = \|\phi(q;\mathcal{W}) - c\|. \tag{8}$$

We impose an explicit regression supervision on the bona fide samples to achieve the optimization goal of Eq. (6) and the implicit metric

[1] The code can be availed from https://github.com/kiran-raja/Residual-MAD/.

learning supervision on the bona fide and morphed images samples to solve Eq. (7) as explained further below.

3.2.1. Learning to classify morphs and bona fide using Negative Log-Likelihood (NLL) Loss

The obtained discriminatory (or pronounced) residuals are used as inputs to a ResNet classifier [44] which produces a score indicating the input image as bona fide or morph. With P_c as input, the classification loss can be formulated as:

$$L_c = \frac{1}{N}\sum_{i=1}^{N} z_i \log q_i + (1 - z_i)\log(1 - q_i), \tag{9}$$

where N is the number of samples, z_i is the binary label and q_i is the network prediction.

3.2.2. Leveraging the imbalance in the dataset through Asymmetric Loss (ASL)

We note that the morphed images and bona fide images can be highly imbalanced corresponding to a real-world scenario. Symmetric loss functions in such a case can help in addressing the asymmetric nature of the dataset as noted in recent work [45]. We, therefore, propose to employ the Asymmetric Loss (ASL) to mitigate the impact of imbalance in training dataset on MAD performance.

Further, asymmetric focusing is shown to reduce the contribution of negative samples to the loss when their probability is low. However, the level of imbalance in bona fide and morph images can be very high in terms of the number of samples and not all face features correspond to morphing, making this attention insufficient. Specifically, when morphing is carried out within the face region keeping the silhouette of one of the contributing subjects as in LMA-UBO [25], face component plays bigger role than silhouette. We, therefore, incorporate the additional asymmetric mechanism of probability shifting to perform hard thresholding of easy negative samples to discard them when their probability is very low [45]. The probability shifting p_m, can therefore be defined as:

$$p_m = \max(p - m, 0) \tag{10}$$

Where the *probability marginm* $\geqslant 0$ is a tunable hyper-parameter.

We further employ two mechanisms of asymmetric focusing and probability shifting in a unified manner motivated by the results reported in earlier works [45], where L_+ and L_- are the positive and negative loss parts, and γ is the focusing parameter:

$$ASL = \begin{cases} L_+ = (1 - p)^{\gamma_+}\log(p) \\ L_- = (p_m)^{\gamma_-}\log(1 - p_m) \end{cases} \tag{11}$$

Where p_m is defined in Eq. (10). ASL allows us to apply two types of asymmetry for reducing the contribution of easy negative samples to the loss function - soft thresholding via the focusing parameters $\gamma_- > \gamma_+$, and hard thresholding via the probability margin m.

It can be convenient to set $\gamma_+ = 0$ so that positive samples will incur simple cross-entropy loss, and control the level of asymmetric focusing via a single hyper-parameter, γ_-. Thus, we employ ASL as a major contributing loss in our proposed approach for classifying morphs, and we refer to this as L_{asl} in the rest of the article.

3.2.3. Auxiliary loss - regression

As the bona fide samples belong to a closed set and the morph samples constitute outliers from this closed set, we impose constraints on the bona fide samples. The morph traces only exist in morphed images, and therefore the morph traces should be close to zero in bona fide samples. It can therefore be safely assumed the zero morph trace as a center of the bona fide in the feature space, and the regression loss on the bona fide images achieves the optimization goal of Eq. (6).

Given a multi-channel image I as input, the residual map of the same size, we hypothesize the residual R to be a zero map for a bona fide sample. It is worth noting that the center C remains unknown for a morphed image owing to various morphing generation processes. The morph trace regression loss for a bona fide sample is the pixel-wise $L1$ loss in the formulation of:

$$L_r = \frac{1}{N_b}\sum_{I_i \in bonafide} ||R_i||_2, \tag{12}$$

where N_b is the number of bona fide in one batch.

3.2.4. Auxiliary loss - triplet loss

We use the metric learning-based loss to promote intra-class compactness in bona fide samples and inter-class separability for bona fide and morphed samples at the feature level. This can be easily seen as an optimization goal of Eq. (7). Specifically, we obtain a set of feature vectors $\{V\}$ by employing the global average pooling (GAP) on feature maps from a layer in each of the ERUs and DRUs and apply the triplet metric supervision in each batch. The triplet metric learning loss can be formulated as:

$$\begin{aligned} L_t &= \frac{1}{T}\textstyle\sum_{i=1}^{T} max(d(a_i,p_i) - d(a_i,n_i) + m, 0), \\ d(i,j) &= \left\| \frac{\mathrm{v}_i}{||\mathrm{v}_i||_2} - \frac{\mathrm{v}_j}{||\mathrm{v}_j||_2} \right\|_2, \end{aligned} \tag{13}$$

where $\{a_i, p_i, n_i\}$ denotes the anchor (bona fide), positive (bona fide), negative (morphed) samples within the ith triplet respectively, T denotes the number of triplets, $d(i,j)$ represents the L2-normalized distance between feature vectors output by the GAP layer, and m is the pre-defined margin constant. We employ online batch-all triplet mining proposed in [46] where at each training step, we collect all the valid triplets within the current batch of data for metric loss computation, the triplets satisfying $||d(a,n) - d(a,p)||_2 < m$.

3.3. Training loss function

The loss functions of the proposed model are fourfold: the binary classification loss L_a, asymmetric loss L_{asl}, the triplet loss L_t and the pixel-wise regression loss L_r for residuals on bona fide samples. The total loss L used for training is thus a combination of all four losses:

Table 1
Details of all five databases with training and testing splits employed in this work.

Generation	Digital				Print-Scan				Print-Scan-Compression			
Type	Bona fide		Morph		Bona fide		Morph		Bona fide		Morph	
	Training	Testing	Trasssining	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
LMA [43]	693	583	1189	1310	693	583	1189	1310	693	583	1189	1310
LMA-UBO [25]	693	583	1203	1318	693	583	1203	1318	693	583	1203	1318
StyleGAN [10]	693	583	1189	1310	693	583	1189	1310	693	583	1189	1310
MIPGAN-I [9]	693	583	1203	1318	693	583	1203	1318	693	583	1203	1318
MIPGAN-II [9]	695	583	1203	1318	695	583	1203	1318	695	583	1203	1318

Table 2
Proposed approach - same set results.

Morphing	Digital			Print Scan			Print Scan Compression		
	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA [43]	0	0	0	0.86	0	0	0.69	0	0
LMA-UBO [25]	2.58	2.06	1.72	7.99	12.15	5.73	9.38	19.97	9.2
StyleGAN [10]	0	0	0	0.08	0	0	0.34	0	0
MIPGAN-I [9]	1.03	0	0	0.34	0	0	0.69	0	0
MIPGAN-II [9]	0.86	0.17	0.17	8.73	29.9	2.58	8.73	31.79	2.92

$$L = \lambda_1 L_c + \lambda_2 L_{asl} + \lambda_3 \sum_{c\in\{ERU-DRU\}} L_t^k + \lambda_4 L_r \tag{14}$$

where k indexes the layer where we apply the triplet loss, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are the regularization weights to balance the influence of the different loss components.

4. Experiments and results

In this section, we present the details of the database employed, the set of experiments conducted and the results obtained to demonstrate the applicability of the proposed approach. We use the dataset presented in earlier works consisting of 5 different types of morphing attacks covering the categories of landmark-based morphs and GAN based morphs. For the convenience of the reader, we present a summary of datasets in this section.

4.1. Databases

All the datasets used in this work are derived from the FRGC-V2 face image database [47] to generate the morph images. The dataset consists of high-quality face images captured from 140 unique subjects (47

Table 3
Ablation study for effectiveness of various losses for MAD classifier trained on LMA-UBO morph types - Digital Domain.

	NLL (L_c)	ASL (L_{asl})	Triplet L_t	Reg-Loss L_r	EER	BPCER_20	BPCER_10
LMA [43]	✓				2.23	1.37	1.37
LMA-UBO [25]	✓				1.2	0.17	0.17
StyleGAN [10]	✓				4.98	4.98	3.26
MIPGAN-1 [9]	✓				28.35	73.2	57.73
MIPGAN-2 [9]	✓				24.66	60.14	47.59
LMA [43]		✓			3.61	3.09	1.72
LMA-UBO [25]		✓			0.91	0.34	0.17
StyleGAN [10]		✓			7.55	13.4	6.53
MIPGAN-1 [9]		✓			15.1	35.74	24.4
MIPGAN-2 [9]		✓			11.84	27.15	14.6
LMA [43]			✓		71.82	97.42	95.88
LMA-UBO [25]			✓		47.95	94.67	91.41
StyleGAN [10]			✓		73.02	97.59	96.22
MIPGAN-1 [9]			✓		65.63	95.19	93.64
MIPGAN-2 [9]			✓		66.54	95.02	93.81
LMA [43]				✓	54.72	97.43	94.17
LMA-UBO [25]				✓	39.10	95.37	88.16
StyleGAN [10]				✓	42.55	93.14	85.42
MIPGAN-1 [9]				✓	38.93	85.42	75.13
MIPGAN-2 [9]				✓	43.80	96.05	89.54
LMA [43]	✓	✓			4.27	2.92	0.86
LMA-UBO [25]	✓	✓			1.37	0.34	0.17
StyleGAN [10]	✓	✓			8.42	11.86	7.39
MIPGAN-1 [9]	✓	✓			12.03	19.24	13.23
MIPGAN-2 [9]	✓	✓			7.28	10.14	5.84
LMA [43]	✓	✓	✓		1.07	0.52	0.52
LMA-UBO [25]	✓	✓	✓		0.86	0.52	0.52
StyleGAN [10]	✓	✓	✓		7.93	23.71	1.37
MIPGAN-1 [9]	✓	✓	✓		35.74	79.9	74.74
MIPGAN-2 [9]	✓	✓	✓		32.3	78.87	70.96
LMA [43]	✓		✓	✓	7.22	9.62	4.98
LMA-UBO [25]	✓		✓	✓	2.88	0.17	0.17
StyleGAN [10]	✓		✓	✓	9.45	14.26	9.11
MIPGAN-1 [9]	✓		✓	✓	19.42	40.55	28.18
MIPGAN-2 [9]	✓		✓	✓	15.02	26.29	20.1
LMA [43]	✓	✓	✓	✓	3.59	2.58	1.72
LMA-UBO [25]	✓	✓	✓	✓	2.58	2.06	1.72
StyleGAN [10]	✓	✓	✓	✓	9.15	16.84	8.59
MIPGAN-1 [9]	✓	✓	✓	✓	12.89	22.51	15.46
MIPGAN-2 [9]	✓	✓	✓	✓	8.93	14.43	8.59

Table 4
Quantitative performance of MAD - Training- Landmarks-I [43]. Results are noted in blue when proposed approach is superior over SOTA or equal to SOTA and noted in red when it is inferior to SOTA.

	Morphing	Digital			Print Scan			Print Scan Compression		
	Generation	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	Ensemble Features [20]	0	0	0	2.35	1.45	0.96	2.58	1.71	1.54
	Hybrid Features [41]	0.16	0	0	1.85	0.85	0.34	2.25	1.12	0.51
	Proposed	0	0	0	0.86	0	0	0.69	0	0
LMA-UBO	Ensemble Features [20]	49.55	92.22	88.85	41.93	81.45	76.25	42.15	83.88	77.64
	Hybrid Features [41]	49.16	99.31	97.59	44.17	86.48	80.24	46.49	88.38	81.95
	Proposed	44.86	82.82	82.82	29.65	99.48	94.85	30.63	97.94	87.97
StyleGAN	Ensemble Features [20]	0.22	0	0	13.36	27.44	16.46	14.77	27.27	19.38
	Hybrid Features [41]	**0.16**	0	0	44.96	83.7	75.47	9.44	14.57	9.14
	Proposed	0.17	0	0	6.36	8.76	4.64	6.03	6.19	3.78
MIPGAN-I	Ensemble Features [20]	39.16	73.14	65.35	9.45	14.57	8.74	8.95	15.26	9.26
	Hybrid Features [41]	46.82	86.62	81.64	12.32	19.72	13.2	9.74	15.95	8.91
	Proposed	16.42	29.21	20.45	33.33	96.05	90.03	25.49	98.11	89
MIPGAN-II	Ensemble Features [20]	34.13	70.49	61.57	5.32	6.68	2.57	6.72	8.16	4.14
	Hybrid Features [41]	44.96	83.7	75.47	5.9	8.42	3.23	5.67	6.18	2.91
	Proposed	13.92	21.82	17.01	8.8	77.66	1.03	5.60	8.06	0.17

female and 93 male) from the FRGC dataset resembling the enrolment passport image quality. A total of 1270 face samples corresponding to 140 data subjects are used in line with the previous works (7–21 samples available for each unique subject).

We employ two morphed datasets created using facial landmarks constrained by Delaunay triangulation with blending [43] (referred to as LMA), landmarks-based techniques with automatic post-processing, and colour equalization (referred to as LMA-UBO) [25]. Further, we choose three different GAN based morphing approaches, which are referred to as StyleGAN [10], MIPGAN-I and MIPGAN-II [9]. All the employed datasets have the images pre-processed to meet the ICAO standards [48] and morphed images with careful selection of subjects based on gender and similarity score using an FRS with realistic and high-quality attacks [43,49].

Further, in line with the previous works, we employ (i) Digital images (bona fide and morphs), (ii) Print-scanned images (bona fide and morphs) and (iii) Print-scanned compressed images (bona fide and morphs). While the digital set consists of bona fide images and morphed images after post-processing to eliminate the artefacts, the print-scanned dataset consists of re-digitized morphed and bona fide images where printing is carried out using a DNP-DS820 [50] in 300 dpi as suggested in ICAO standards [48]. Print-scanned compressed images consist of morphed and bona fide images compressed to have a size of 15kB, making them suitable to store in the e-passport. Both Print-scanned and Print-scanned compressed subsets mimic real-world settings of the passport application and issuance processes.

The reader is further referred to the original works to get complete details of dataset [10,9,43], however, we present the details on the

Table 5
Quantitative performance of MAD trained on LMA-UBO data [25]. Results are noted in blue when proposed approach is superior over SOTA or equal to SOTA and noted in red when it is inferior to SOTA.

	Morphing	Digital			Print Scan			Print Scan Compression		
	Generation	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	Ensemble Features [20]	48.57	97.77	95.36	24.19	52.48	43.22	21.64	47.51	36.19
	Hybrid Features [41]	45.67	96.91	94.16	32.26	77.87	66.55	24.51	50.94	40.65
	Proposed	3.59	2.58	1.72	9.85	19.97	9.72	13.82	44.44	21.01
LMA-UBO	Ensemble Features [20]	3.62	2.22	0.68	6.32	7.97	2.42	5.57	6.41	2.42
	Hybrid Features [41]	1.53	0.17	0	5.21	5.19	3.14	5.37	5.71	3.46
	Proposed	2.58	2.06	1.72	7.99	12.15	5.73	9.38	19.97	9.2
StyleGAN	Ensemble Features [20]	29.67	61.92	52.48	27.18	61.57	50.6	29.18	62.14	52.48
	Hybrid Features [41]	34.76	74.44	62.95	34.8	67.23	58.14	23.17	49.22	38.25
	Proposed	9.15	16.84	8.59	8.16	13.19	6.25	8.09	11.81	6.25
MIPGAN-I	Ensemble Features [20]	30.23	65.35	53.17	43.92	87.65	79.24	44.24	89.23	82.33
	Hybrid Features [41]	46.29	84.04	77.01	34.16	71.18	64.66	35.5	76.84	65.52
	Proposed	12.89	22.51	15.46	6.45	7.81	3.82	1.44	0.52	0.17
MIPGAN-II	Ensemble Features [20]	27.13	58.83	45.45	33.57	77.35	65.52	40.46	84.9	75.47
	Hybrid Features [41]	46.82	83.53	75.81	35.91	77.18	65.24	36.5	79.24	68.78
	Proposed	8.93	14.43	8.59	4.51	4.51	2.6	2.26	1.04	0.17

Table 6
Quantitative performance of proposed MAD trained with LMA-UBO [25] with increased number of epochs (60).

Morphing	Digital			Print Scan			Print Scan Compression		
	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	3.61	2.06	0.34	21.07	81.77	71.01	30.31	88.37	75.69
LMA-UBO	0.68	0.17	0.17	7.22	33.16	0.35	7.45	18.23	1.22
StyleGAN	13.73	31.1	19.07	16.32	73.09	55.73	11.52	51.22	21.88
MIPGAN-I	25.43	58.93	47.94	4.17	0.87	0.17	5.08	5.21	1.39
MIPGAN-II	24.73	58.59	46.91	2.05	0.17	0.17	2.05	0.52	0.35

total number of images used in this work in Table 1. The attack strength of the databases in the form of vulnerability analysis of FRS is further provided in Tables A.1, A.2, A.3.

4.2. Protocols

We adopt the evaluation protocols as described in earlier works [9,20,40] to evaluate the proposed MAD algorithm by dividing the dataset into training, validation and testing set that consists of independent data subjects with no overlap between the splits. In a slight modification to the recent protocol, we derive a validation set from the training set to validate the learning of the network where 30% of the training set is used as the validation set. Similar to earlier works, we provide within database (training and testing dataset from the same morph generation approach) and cross-database-evaluation of MAD mechanisms on digital, print-scan and print-scan with compression data types.

All the results obtained on the proposed S-MAD on five different types of generation mechanisms are reported using the ISO/IEC metrics [51] which are specifically measuring the detection accuracy of attacks, namely APCER (%) and BPCER (%), along with the EER (%).

4.3. Training details

We train the proposed approach by fixing the number of epochs to 30, a learning rate of $5e-4$, with a batch size of 32 for all the experiments. Further, we present the individual analysis of hyper-parameters through empirical validation, discussing the impact of each of the different loss functions in Section 4.6. The training and testing are conducted on an Nvidia 2080 Ti GPU enabled computer with a Linux operating system (Ubuntu 20.04).

4.4. SOTA benchmarks

Of the number of works available from state-of-the-art MAD [25,27,31,52–56], we choose to compare our results to a recent benchmark provided on the datasets used in our work [9]. Specifically, we benchmark our approach against Hybrid features [41] and Ensemble features [20] for detecting morphing attacks based on the performance obtained in NIST FRVT MORPH challenge [33] with the best performance in detecting printed and scanned morph images. While the Hybrid features [41] use both scale space and colour space combined with multiple classifiers, the Ensemble features [20] use textural features in conjunction with a set of classifiers.

4.5. Results - known set testing (closed-set)

We first establish the applicability of the proposed approach using the known training and testing set in a closed-set protocol. The key motivation for this evaluation is to validate the applicability when the testing set resembles the training set characteristics. The results are presented in Table 2 where one can note less then 3% EER and BPCER for digital dataset. The proposed approach also obtains near ideal error rates for most cases in print-scan and print-scan-compression except for LMA-UBO [25] and MIPGAN-II [9] sets. The results suggest promising nature of proposed approach for 11 of 15 different cases.

Table 7
Quantitative performance of MAD trained on MIPGAN-I data [9]. Results are noted in blue when proposed approach is superior over SOTA or equal to SOTA and noted in red when it is inferior to SOTA.

	Morphing	Digital			Print Scan			Print Scan Compression		
	Generation	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	Ensemble Features [20]	23.66	51.45	39.96	5.82	7.22	2.92	6.17	7.54	3.94
	Hybrid Features [41]	47.15	87.16	79.41	6.5	8.23	4.15	7.91	10.29	6.34
	Proposed	18.7	31.62	25.43	4.12	3.95	1.89	8.02	11.51	7.39
LMA-UBO	Ensemble Features [20]	35.38	82.33	68.95	41.67	95.14	83.53	43.68	96.01	85.44
	Hybrid Features [41]	28.62	75.64	61.4	44.38	95.66	85.78	38.18	90.46	78.16
	Proposed	57.22	90.21	85.57	32.08	82.82	71.31	33.51	77.66	68.73
StyleGAN	Ensemble Features [20]	17.72	37.22	26.58	12.19	26.24	15.26	11.82	24.69	14.23
	Hybrid Features [41]	31.16	64.32	53.85	11.99	**19.2**	13.72	9.93	18.15	9.94
	Proposed	10.14	15.12	10.14	11.37	19.24	12.54	10.3	19.93	10.65
MIPGAN-I	Ensemble Features [20]	0	0	0	0	0	0	0	0	0
	Hybrid Features [41]	0	0	0	0	0	0	0	0	0
	Proposed	1.03	0	0	0.34	0	0	0.69	0	0
MIPGAN-II	Ensemble Features [20]	2.15	0.17	0	0.68	0	0	0.64	0	0
	Hybrid Features [41]	1.36	0.34	0	0.86	0	0	0.84	0	0
	Proposed	0.03	0	0	0.3	0	0	0.53	0	0

4.6. Choosing weights through ablation studies

The results presented in Table 2 are based on the chosen weights for different loss combinations. Thus, we first study the impact of various losses before evaluating open-set protocol for generalizability. The impact of various losses can be seen from Table 3 when trained on LMA-UBO and tested on different digital morph data. While the NLL loss provides very high detection accuracy for LMA, LMA-UBO and StyleGAN, the same deteriorate for MIPGAN-I and MIPGAN-II. On the other hand, the ASL has better scalability for MIPGAN-I and MIPGAN-II while losing some performance in LMA and StyleGAN. However, Triplet Loss and Registration Loss do not contribute heavily to improving the performance. Combining all four losses provides a stable performance across the same-set and cross-set scenarios. Specifically, using greater weights on NLL and ASL with lower weights on Triplet and Registration provides a balanced but not ideal performance. We have chosen to employ the weights as $\lambda_1 = 0.6, \lambda_1 = 0.2, \lambda_1 = 0.15$ and $\lambda_1 = 0.05$ for all the experiments reported further below. It should, however, be noted that the performance can be further tuned by optimizing the weights for loss function based on the availability of data. The impact of chosen weights is evident when all the loss functions are incorporated in the training where EER, BPCER_20 and BPCER_10 result in lower errors than either NLL, ASL or a combination of them.

4.7. Results - unknown set testing (open-set)

We extend the experiments to verify the applicability of the proposed approach as a generalizable MAD algorithm by benchmarking it against two NIST evaluated MAD algorithms. For such an evaluation, we train the proposed approach using the data from LMA, LMA-UBO, StyleGAN, MIPGAN-I and MIPGAN-II independently in a corresponding setting of digital, print-scan and print-scan-compression. While we present results from LMA, LMA-UBO, and MIPGAN-I here, we present results and analysis from StyleGAN and MIPGAN-II in Appendix (Section E and Section D respectively), due to page constraints.

4.7.1. Unknown testing - LMA trained

Table 4 presents the results obtained using the proposed approach in an unknown testing set scenario. All the results are compared against two SOTA approaches and we note the following observations accordingly:

- The proposed approach obtained significantly lower EER in 9 out of 12 cross-data settings performing better than any of the two SOTA approaches based on Hybrid features [41] and Ensemble features [20]. Along with the EER, we note a lower BPCER_10 in 10 of 12 cases. A lower error rate of BPCER_20 and BPCER_10 further indicates

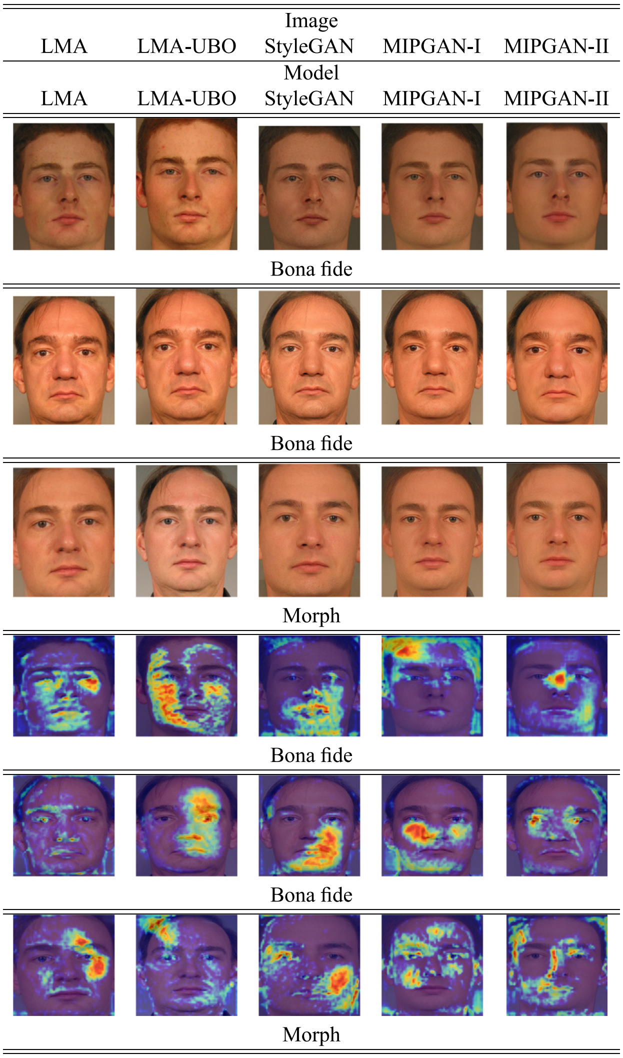


Fig. 4. XGradCAM analysis of proposed approach on bona fide and morphed images in digital domain.

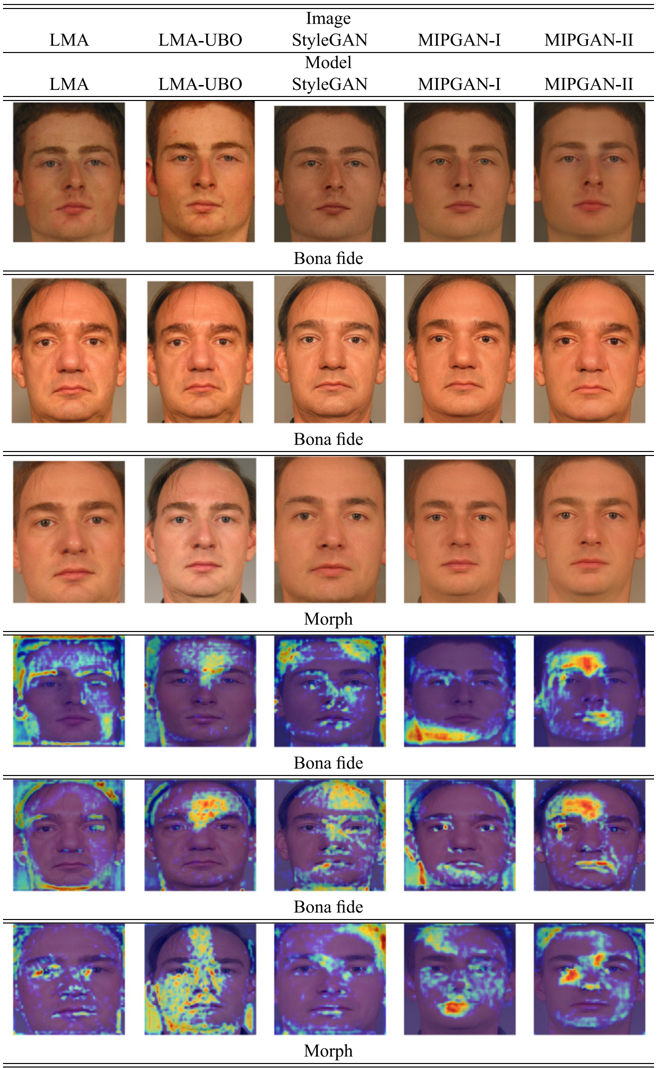


Fig. 5. AblationCAM analysis of proposed approach on bona fide and morphed images in digital domain.

the superior performance of the proposed approach in detecting morphing attacks.

- We note that the proposed approach can perform better for styleGAN, MIPGAN-I and MIPGAN-II in the digital domain, while it degrades for MIPGAN-I after print-scan and print-scan-compression. Although a lower performance can be observed for MIPGAN-II under print-scan, the performance is comparable for MIPGAN-II in the print-scan-compression setting.
- Intrigued by the low performance of the proposed approach on LMA-UBO generally, we analyze the scores and note that the scores tend to be very biased towards bona fide or morphs leading to high BPCER. Due to such binning of scores, it is natural that the BPCER metric, which is obtained at fixed APCER, tends to be very high (Refer Fig. C.13 in Appendix).

4.7.2. Unknown testing - LMA-UBO trained

In line with previous experiments, we train the proposed approach using LMA-UBO data, whose results are presented in Table 5. While we note a superior performance when trained with LMA-UBO and tested on different datasets, the known set testing suffers from performance degradation. We note specific observations from this set of experiments:

- The proposed approach reduces the EER in all unknown testing set scenarios. The EER in the case when LMA is tested equals 3.59%, reducing the error rate from 45.67% in the digital domain.
- At the same time, reduction in error rates is also reduced significantly from 24.19% to 9,85% and 21.64% to 13.82% for print-scan and print-scan-compression respectively for LMA as testing set.
- The proposed approach reduces the error rates significantly for StyleGAN, MIPGAN-I and MIPGAN-II. The EER is observed to be 9.15%, 8.16% and 8.09% for StyleGAN for digital, print-scan and print-scan-compression while simultaneously reducing BPCER at APCER = 5% and APCER = 10%.
- Similar to StyleGAN, MIPGAN-I and MIPGAN-II also is detected with high accuracy with 12.89%, 6.45% and 1.44% for digital, print-scan and print-scan-compression, respectively in MIPGAN-I. Similarly, MIPGAN-II attacks are detected with high accuracy with 8.93%, 4.51% and 2.26% for digital, print-scan and print-scan-compression, respectively.
- We investigate it further by increasing the number of training epochs to verify if the robustness of the proposed MAD approach increases. As noted from the results in Table 6, the results for the closed-set (LMA-UBO) and open-set MIPGAN-I and MIPGAN-II decreases, and the error rates for LMA and StyleGAN in print-scan and print-scan-compression

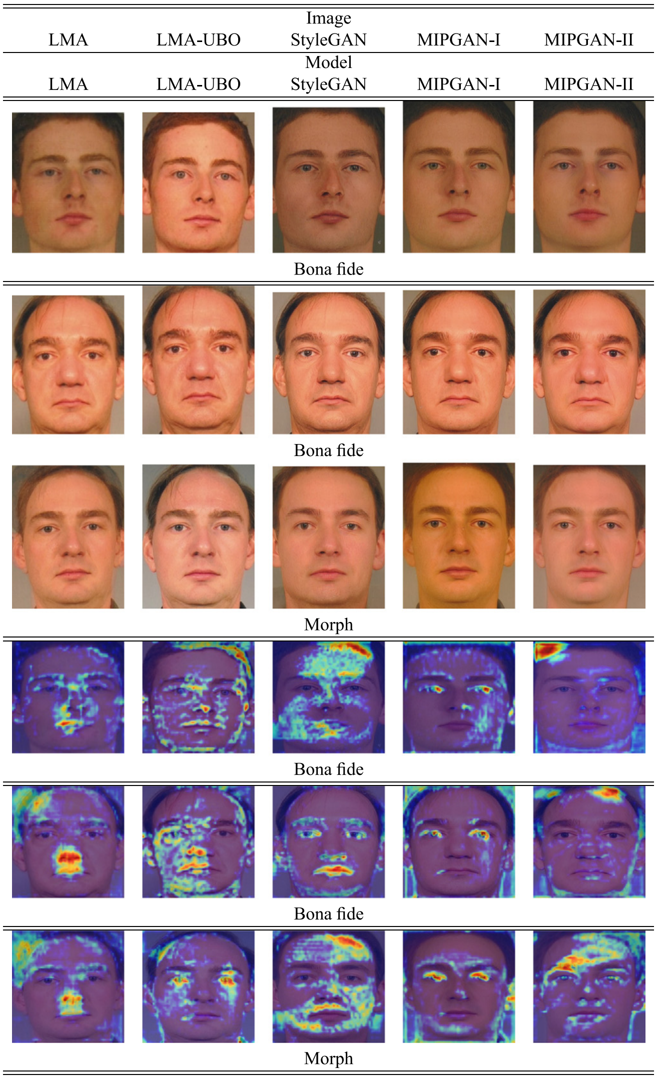


Fig. 6. AblationCAM analysis of proposed approach on bona fide and morphed images in print-scan domain.

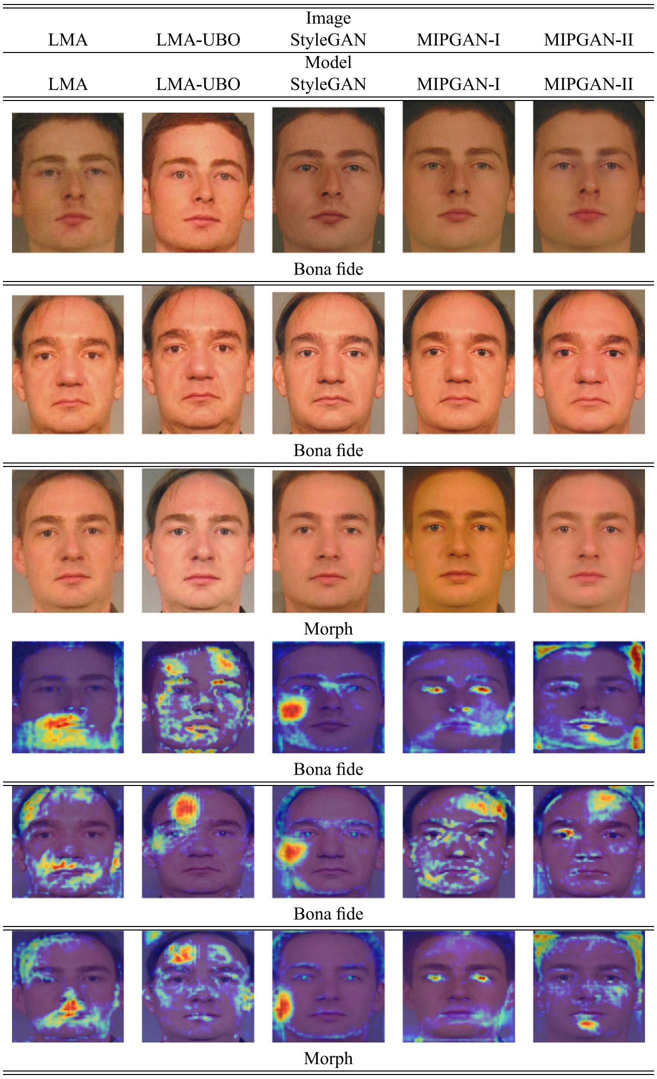


Fig. 7. AblationCAM analysis of proposed approach on bona fide and morphed images in print-scan-compression domain.

increase. Such an observation can be argued from over-fitting parameters favouring one kind of data. However, we hypothesise that such a challenge can be mitigated by employing various morph generation algorithms in training in future works.

4.7.3. Unknown testing - MIPGAN-I trained

We analyze the results obtained using MIPGAN-I data in the training set, and the same is presented in Table 7. Based on the results obtained, one can make the following observations:

- The proposed approach, when trained with MIPGAN-I, obtains lower EER rates in 9 of the 12 individual cases in unknown data testing.
- MIPGAN-I training detects MIPGAN-II attacks in an unknown setting but fails to detect LMO-UBO as the unknown data. This observation indicates the limited generalization of proposed MAD when trained on GAN data and tested on landmark-based data.
- The proposed approach slightly deteriorates against SOTA approaches when print-scan-compression data from the StyleGAN approach is presented. As the data StyleGAN generation is significantly different to MIPGAN-I data, this degradation can be addressed by incorporating the StyleGAN data into the training set. However, we refrain from this as the focus of the work is to study the generalizability of the proposed approach when the testing data is completely unknown.

5. Explainability and discussion

To better understand the proposed MAD approach, we analyse images using the Class Activation Mapping (CAM) on all the trained models. We first analyse the normalised activations by using GradCAM with scaled gradients [57]. GradCAM analysis illustrates the coarse localisation map highlighting important regions in the image for predicting the ground truth, and we employ the same to visualise the regions employed for morph classification in the proposed approach. We, therefore, employ images from two bona fide face images from the testing set and the corresponding morph as illustrated in Fig. 4 to conduct this set of analyses. For simplicity, we demonstrate the activation maps obtained from the *conv* layer of the decoder block in our proposed approach and a similar analysis can be extended for other layers.

Some key observations from this analysis correspond to the morph generation process to a greater degree, as noted below:

- As noted in Fig. 4, the first column represents the bona fide and morphs from LMA [43], and in this case, the proposed approach

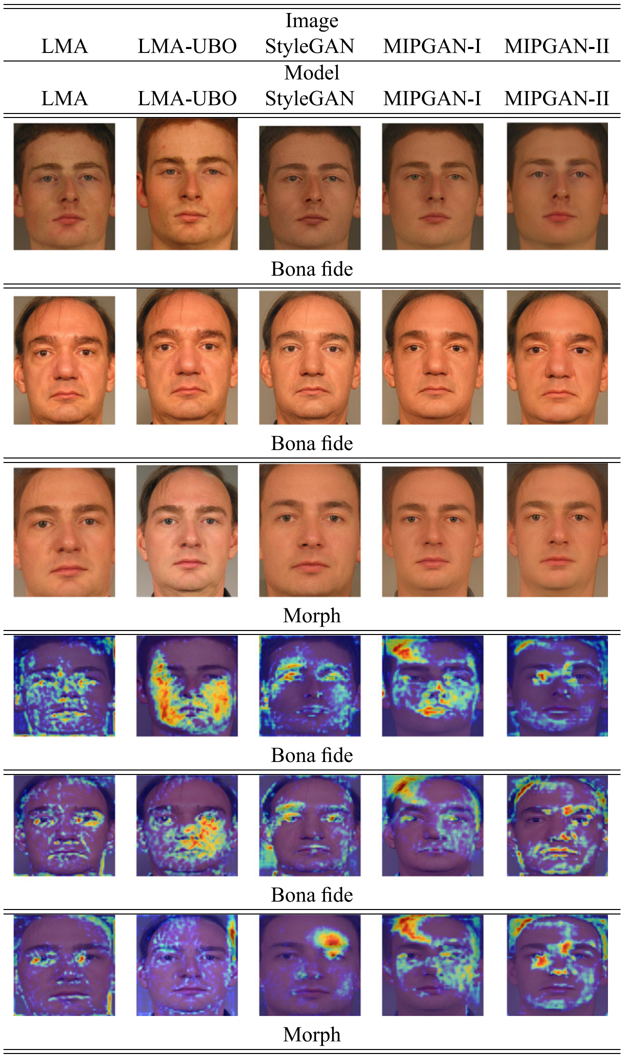


Fig. 8. ScoreCAM analysis of proposed approach on bona fide and morphed images in digital domain.

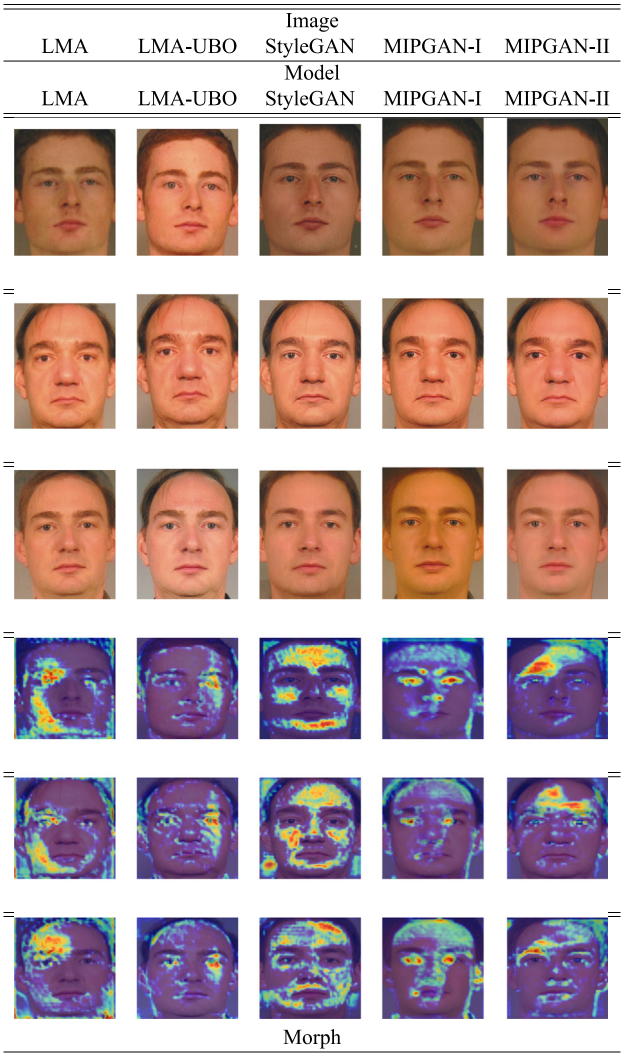


Fig. 9. ScoreCAM analysis of proposed approach on bona fide and morphed images in print-scan domain.

focuses on the lip region and eyebrow region for detecting morphs. The LMA approach [43] employs landmarks and the resulting artefacts around the lip and the eyebrow region has been processed in this dataset. Our proposed approach focuses on these regions to detect morphs.

- The second column illustrates the activation maps of LMA-UBO [25] where the morphs are processed along with the silhouette based on the highest contributing subject chosen for morphing, and the same can be seen in the last row of the second column. The silhouette of the face region appears to be activated largely from the proposed approach, along with the regions around the eye.
- Further, StyleGAN [10] based morphing illustrated in the third column does not employ any landmarks, and the same can be observed in images where the network focuses on the regions inside the face area.
- While in MIPGAN-I [9] shown in column 4, the proposed approach tends to focus on regions around the face as the process of morph generation itself is a superior version of StyleGAN with identity priors enforced, making the identity information stronger. It is also interesting to note that the proposed approach in such a case tends to focus on the eye region for detecting the morphs. However, it should also be carefully noted that activations take into consideration the hair region where visible artefacts can be seen due MIPGAN-I generation process [9]. A similar observation can be drawn for the proposed approach trained within MIPGAN-II [9] which is a variant of MIPGAN-I.

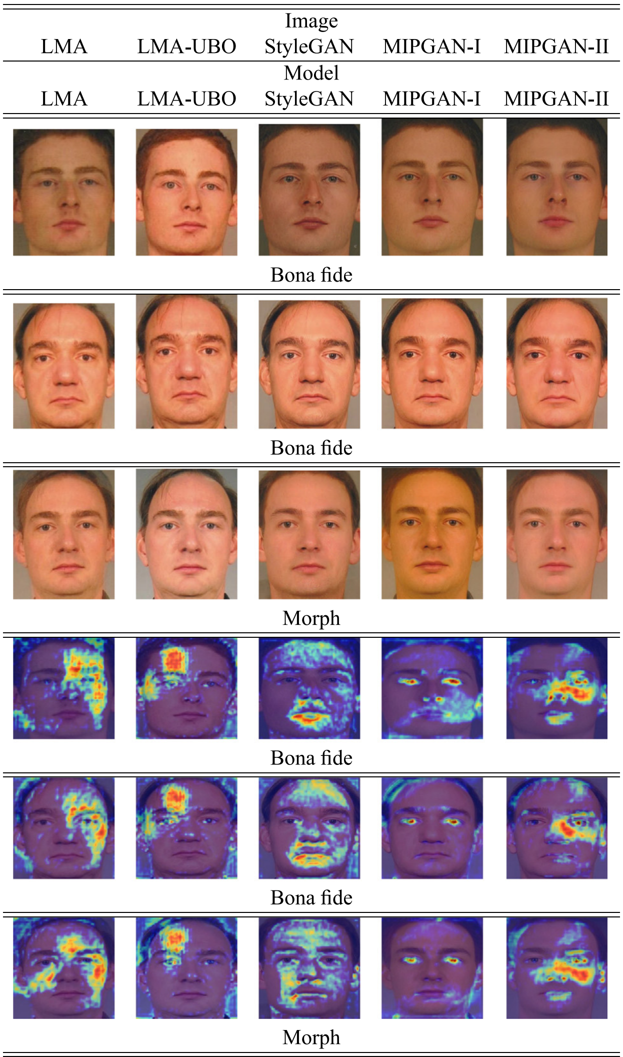


Fig. 10. ScoreCAM analysis of proposed approach on bona fide and morphed images in print-scan-compression domain.

Noting the Fig. 4, it can be deduced that the proposed approach can be made more effective by carefully cropping the images in the pre-processing step such that the network can further uniformly focus on important regions to achieve better accuracy.

5.1. Complementary CAM analysis

We further employ Ablation-based CAM (AblationCAM) [58] and ScoreCAM [59]. AblationCAM uses ablation analysis to determine the importance (weights) of individual feature map units w.r.t. class and generate gradient-free visual explanations for the proposed approach. AblationCAM can be used to produce a coarse localisation map highlighting the important regions in the image for predicting the concept [58]. On the other hand, Score-CAM provides a gradient-free visual explanation bridging the gap between perturbation-based and CAM-based methods and intuitively representing the weight of activation maps. Score-CAM incorporates network confidence in deriving weight for each activation map. While AblationCAM can help understand the important regions through a class-discriminative approach, ScoreCAM obtains the weight of each activation map through its forward passing score on the target class resulting in a linear combination of weights and activation maps. Thus, these approaches can provide complementary visualisation of the proposed MAD.

5.1.1. Analysis from AblationCAM

We analyze the results for the proposed approach for digital, print-scan and print-scan-compression images (bona fide and morph) as presented in Figs. 5–7, respectively. As noted from Fig. 5, the proposed approach focuses on the regions corresponding to facial silhouette and landmark areas, including the areas such as the chin, eyebrows and eye region. It can be further noted that the activations of high intensity above the forehead (between forehead and transition to hair) where traces of morphing can be seen, despite post-processing. Nonetheless, the network occasionally seeks the background area in making the decisions and this can be easily argued from the nature of data as seen in Fig. 5 for digital images.

In the case of Figs. 6 and 7 corresponding to print-scan and print-scan compression, high activation can be observed for models trained on LMA, LMA-UBO. StyleGAN and MIPGAN-I, while incorrect activations can be noted for MIPGAN-II. This observation can be easily correlated to the low performance noted in Table E.5 where the clear failure of MIPGAN-II trained models in generalisation can be observed.

5.1.2. Analysis from ScoreCAM

We analyze the results from the proposed approach for digital, print-scan and print-scan-compression images (bona fide and morph) as presented in Figs. 8–10, respectively. It can be noted that weights of each activation map through its forward passing score on target class results in predictions close to the ground-truth (i.e., morphs and bona fide). The results indicate the decisions explainable with respect to the morphing generation type as noted from Figs. 8–10.

5.1.3. Interpreting CAM maps and uncertainty in explainability

CAM maps illustrated in Figs. 5–7, can help the human observer to make a decision. Specifically, a higher activation, i.e., most intense (red) areas, provide the regions for deciding an image as bona fide or morph. The activations around the face silhouette in bona fide images, as shown in Fig. 11 indicate the information used in determining an image as bona fide, and this appears to be relatively less intense for a

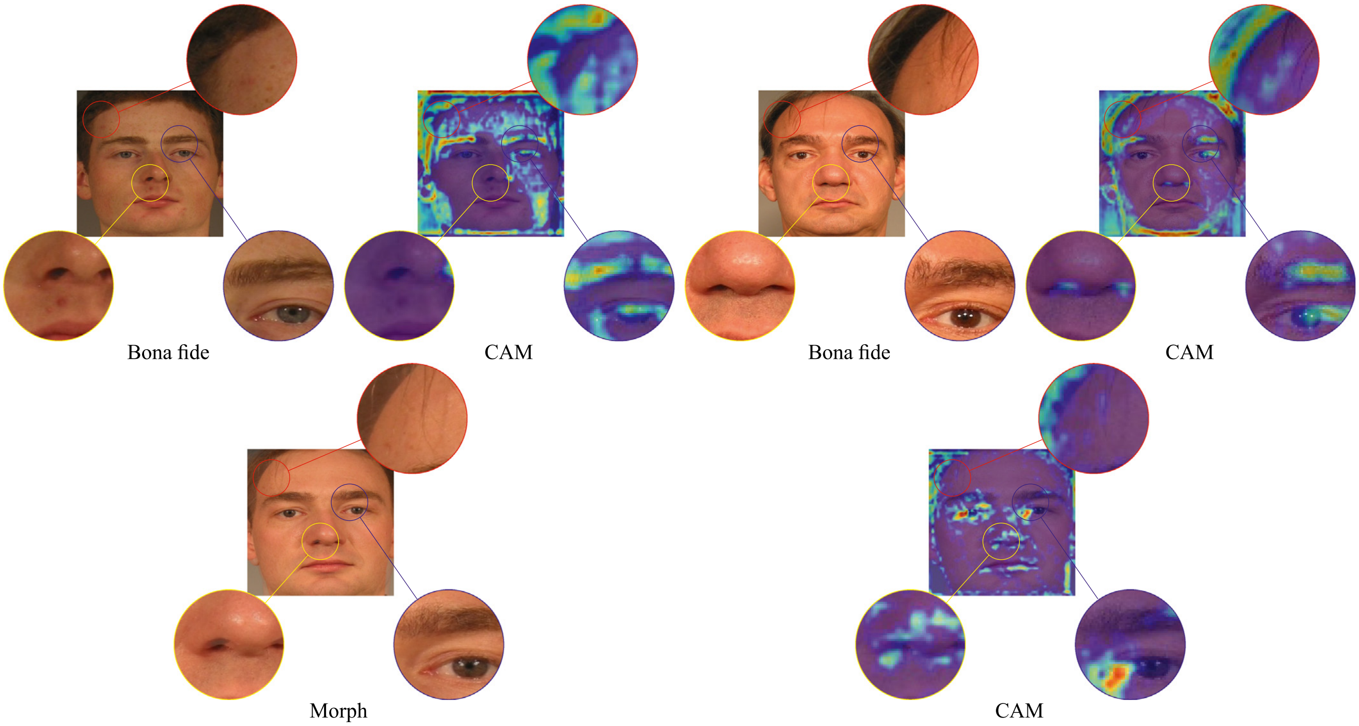


Fig. 11. Uncertainty in activation maps from AblationCAM analysis of proposed approach on bona fide and morphed images in digital domain for LMA.

morphed image. For the model trained using LMA data, intense activation around the eye region and silhouette can be used for deciding an image as bona fide, while the heavy activations around nostrils can indicate the probability of morphing and post-processing as illustrated in Fig. 11.

While the activation maps provide a reasonable way of interpreting the decision, we also illustrate that activation maps do not take the model and data uncertainties into account. Fig. 11 presents an illustration of such potential uncertainties for a model trained on LMA data. As it can be noticed, the activation is not uniform for both

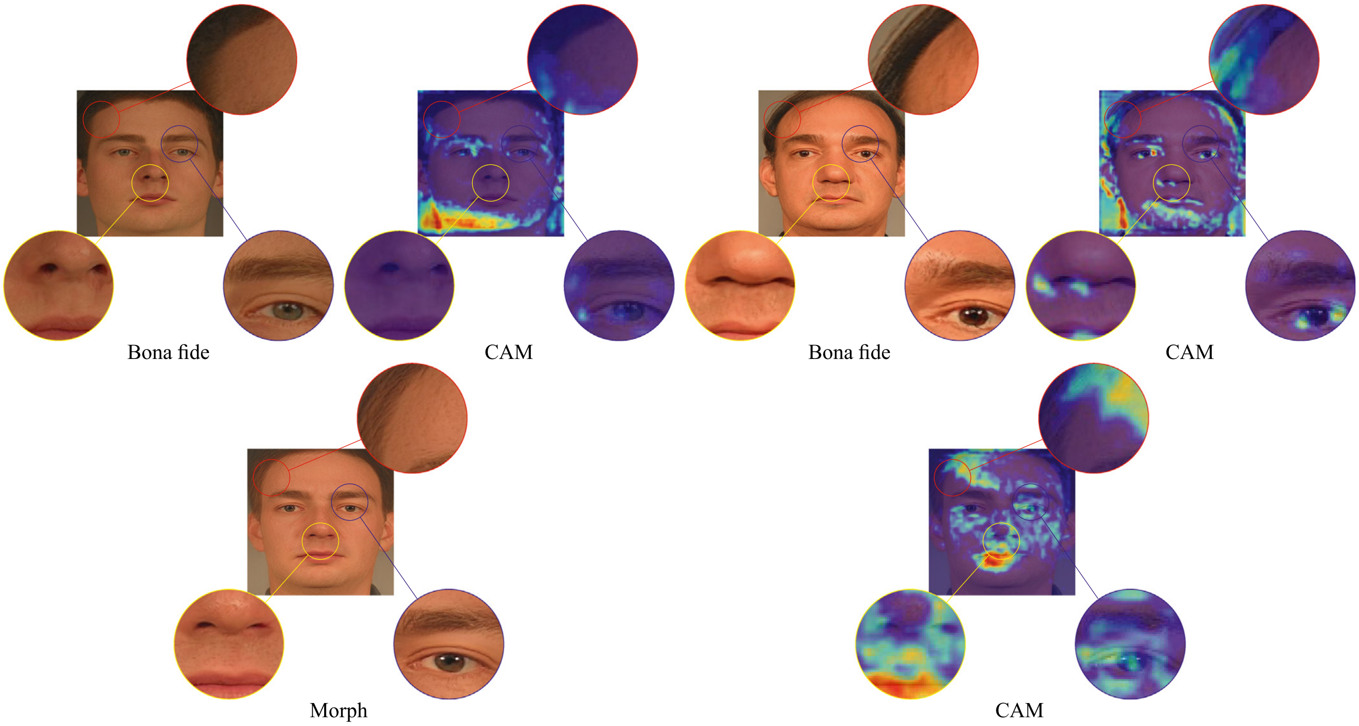


Fig. 12. Uncertainty in activation maps from AblationCAM analysis of proposed approach on bona fide and morphed images in digital domain for MIPGAN-I.

bona fide and morphed images. Looking into the zoomed areas of the eye region and nostrils, one can note that the activation is different for bona fide and morph. Such differences in activation maps can help determine the morphs when a human observer looks at the activation maps. However, the observations are inverted for the model trained on MIPGAN-I for the same set of images as shown in Fig. 12. Thus, a human observer looking at CAM maps as an explainability tool should be aware of model variations and the uncertainties.

6. Conclusion

MAD algorithms need to be robust enough to detect the morphed images of unknown generation types to be usable in operational settings. Noting the limited performance of MAD algorithms in detecting unknown morph types in testing, we have proposed a new approach in this work that generalizes better compared to SOTA algorithms. The proposed approach uses a multi-stage encoder-decoder network to learn the residuals across different colour spaces to detect morphing attacks. The proposed approach achieves a generalizable MAD by learning a linear classifier with cross-entropy loss, asymmetric loss, regression and triplet loss We have conducted experiments on five different datasets created using landmark-based morphs and GAN-based morphs where images are available in the digital domain, print-scan and print-scan-compression domain. The obtained results indicate a near-ideal performance of the proposed MAD with an Equal Error Rate (EER) of 0% in the best case and 2.58% in the worst case in the digital domain. The applicability of the proposed approach to 60 different combinations is illustrated where the testing set consists of data from unknown morphing generation to study the generalization ability of our proposed approach. By training the proposed approach on landmark-based morph generation data, we obtain an EER of 3.59% in the best case and 12.89% in the worst case for morphed images in the digital domain, reducing the error rates from 45.67% and 30.23% respectively. The analysis for explainability is further presented to analyze the decisions using three different CAM analyses. Future works in this direction should investigate reducing the total number of channels by analyzing channel-wise importance to improve generalizability further.

CRediT authorship contribution statement

Kiran Raja: Conceptualization, Methodology, Software, Writing – original draft. **Gourav Gupta:** Data curation, Investigation. **Sushma Venkatesh:** Resources, Methodology, Writing – review & editing. **Raghavendra Ramachandra:** Resources, Methodology, Writing – review & editing. **Christoph Busch:** Methodology, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kiran Raja reports financial support was provided by European Union.

Appendix A. Attack potential of the databases

We present the attack potential of the databases by studying the vulnerability of FRS using Mated Morphed Presentation Match Rate (MMPMR) [49] and Fully Mated Morphed Presentation Match Rate (FMMPMR) [10] based on the threshold provided by respective FRS. We report the vulnerability of COTS FRS - Cognitec-FRS [60] and Neurotech [61] along with an open source FRS -Arcface [62]. Further, to effectively analyse the vulnerability, we also present the results using Relative Morph Match Rate (RMMR) [4].RMMR can be related to MMPMR and FMMPMR as given by Eqs. (A.1) and (A.2).

$$RMMR(\tau)_{MMPMR} = 1 + (MMPMR(\tau)) - [1 - FNMR(\tau)] \tag{A.1}$$

$$RMMR(\tau)_{FMMPMR} = 1 + (FMMPMR(\tau)) - [1 - FNMR(\tau)] \tag{A.2}$$

where, F NMR indicates the False Reject Rate (F NMR) of the FRS under consideration obtained at the threshold τ. In this work, τ represents the value corresponding to $FMR = 0.1\%$ compliant to FRONTEX FAR/FRR constraints. We present F NMR corresponding to the FRS to calculate the RMMR. It has to be noted that RMMR in Eqs. (A.1) and (A.2) equals to MMPMR/FMMPMR when F NMR = 0.

A detailed analysis of the MMPMR of these datasets according to gender distribution can be obtained in corresponding articles - LMA [43], LMA-UBO [25], StyleGAN [10], MIPGAN-I [9] and MIPGAN-II [9].

Table A.1
Vulnerability of COTS Cognitec-FRS [60] for various morph generation approaches. As $FNMR = 0@FMR = 0.1\%$ for Cognitec-FRS [60] following Eqs. (A.1) and (A.2), the value of RMMR is equal to MMPMR/FMMPMR.

Morph generation	MMPMR/RMMR (%)	FMMPMR/RMMR (%)	MMPMR/RMMR (%)	FMMPMR/RMMR (%)	MMPMR/RMMR (%)	FMMPMR/RMMR (%)
	Digital		Print-Scan		Print-Scan with compression	
Landmark-I [43]	**100**	**98.84**	**97.64**	**97.60**	**97.84**	**97.30**
Landmark-II [25]	88.65	78.72	91.85	81.56	90.61	79.33
StyleGAN [10]	64.68	41.49	61.72	39.90	58.92	35.89
MIPGAN-I [9]	94.36	84.65	92.97	82.23	92.29	79.88
MIPGAN-II [9]	92.93	81.59	80.56	79.02	90.24	75.20

Table A.2
Vulnerability of COTS Neurotech [61] for various morph generation approaches. As $FNMR = 0@FMR = 0.1\%$ for Neurotech [61] following Eqs. (A.1) and (A.2), the value of RMMR is equal to MMPMR/FMMPMR.

Morph generation	MMPMR/RMMR (%)	FMMPMR/RMMR (%)	MMPMR/RMMR (%)	FMMPMR/RMMR(%)	MMPMR/RMMR(%)	FMMPMR/RMMR(%)
	Digital		Print-Scan		Print-Scan with compression	
Landmark-I [43]	**99.51**	**95.37**	**96.32**	**85.43**	**94.30**	**79.25**
Landmark-II [25]	90.16	71.17	90.59	66.67	83.50	57.38
StyleGAN [10]	55.06	29.39	36.36	14.83	35.62	14.28
MIPGAN-I [9]	63.22	35.73	40.46	28.71	61.66	34.14
MIPGAN-II [9]	57.47	31.45	51.72	23.54	54.94	27.46

Table A.3
Vulnerability of Arcface [62] FRS for various morph generation approaches. $FNMR = 0@FMR = 0.1\%$ for Arcface [62] following Eqs. (A.1) and (A.2), the value of RMMR is equal to MMPMR/FMMPMR.

Morph generation	MMPMR/RMMR(%)	FMMPMR/RMMR(%)	MMPMR/RMMR(%)	FMMPMR/RMMR(%)	MMPMR/RMMR(%)	FMMPMR/RMMR(%)
	Digital		Print-Scan		Print-Scan with compression	
Landmark-I [43]	**99.68**	**98.00**	**97.88**	**96.89**	**97.84**	**96.75**
Landmark-II [25]	91.79	84.96	94.33	86.96	94.53	86.54
StyleGAN [10]	72.80	56.95	75.60	59.79	75.16	59.51
MIPGAN-I [9]	94.45	85.94	93.81	85.46	93.97	85.48
MIPGAN-II [9]	94.21	86.94	94.05	85.95	93.85	85.77

Appendix B. Training details

We train the proposed approach by fixing the number of epochs to 30, a learning rate of $5e-4$, with a batch size of 32 for all the experiments. Further, we present the individual analysis of hyper-parameters through empirical validation, discussing the impact of each of the different loss functions in Section 4.6. All the training and testing is conducted on a Nvidia 2080 Ti GPU enabled computer with Linux operating system.

Appendix C. Overlap of scores leading to high BPCER in MAD

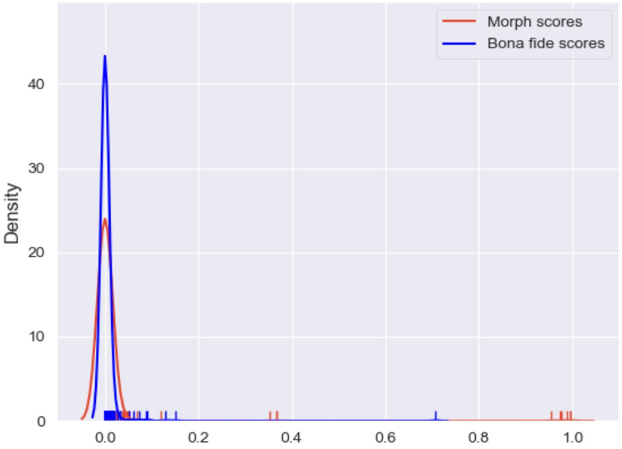


Fig. C.13. Overlap of scores leading to high overlap of bona fide and morph scores when trained with LMA data and tested on LMA-UBO data.

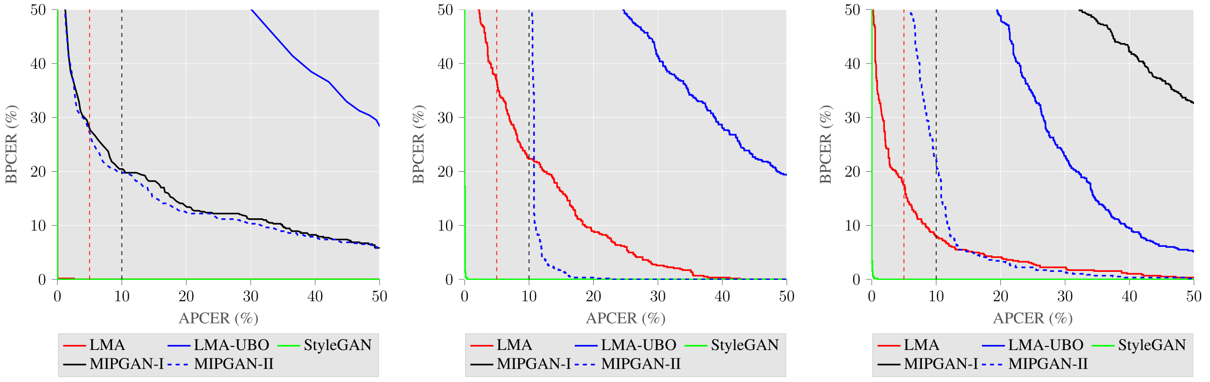
.

Appendix D. Unknown testing - StyleGAN trained

We further train the proposed MAD using StyleGAN data, and the results are noted in Table D.4. Unlike the previous two results presented, the MAD trained on StyleGAN data performs relatively poor, indicating limited generalizability. We note our observations as presented below:

- The proposed approach obtains lower EER rates in 9 of the 12 individual cases in unknown data testing.
- Despite the obtained performance, we note very high EER and BPCER in LMA-UBO, indicating low generalizability of the proposed approach when trained on StyleGAN data irrespective of digital, print-scan and print-scan-compression.
- A similar drop in performance can also be noted for MIPGAN-I and MIPGAN-II data in print-scan and print-scan-compression, for which further analysis is provided below.

We further analyse the high error rates using the Detection Error Trade-off (DET) curves to understand the proposed approach's low generalisation when trained with the StyleGAN data. As noted from Fig. D.14 and Table D.4, the proposed approach performs poorly for MIPGAN-I and MIPGAN-II. However, a closer inspection of the DET curves for print-scan and print-scan-compression reveals low and near-ideal BPCER at APCER = 20%. The analysis demonstrates the grouping of scores, and due to the nature of scores which are highly dense around 0 and 1, the BPCER_20 is observed to be very high. Future work in direction should also investigate the widening of the score range to make the approach robust when trained with StyleGAN data.



(a) Digital (b) Print-Scan (c) Print-Scan-Compression

Fig. D.14. DET curves obtained on proposed approach with StyleGAN data in training.

Table D.4
Quantitative performance of MAD trained on StyleGAN data [10]. Results are noted in blue when proposed approach is superior over SOTA or equal to SOTA and noted in red when it is inferior to SOTA. (*Refer Fig. D.14 for further analysis of the obtained performance for models trained with StyleGAN data.)

	Morphing	Digital			Print Scan			Print Scan Compression		
	Generation	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	Ensemble Features [20]	0.32	0	0	16.6	28.13	19.89	13.89	22.12	17.66
	Hybrid Features [41]	0.42	0	0	15.26	26.41	17.66	14.37	22.81	16.92
	Proposed	0.17	0	0	15.02	36.08	22.34	8.93	17.35	7.90
LMA-UBO	Ensemble Features [20]	44.72	89.53	80.61	38.31	78.5	69.15	38.84	83.7	74.17
	Hybrid Features [41]	45.65	90.22	84.56	34.18	81.95	70.53	32.93	78.5	64.12
	Proposed	38.94	74.4	58.25	34.72	99.66	93.13	27.32	92.10	83.85
StyleGAN	Ensemble Features [20]	0	0	0	0	0	0	0	0	0
	Hybrid Features [41]	0	0	0	0	0	0	0	0	0
	Proposed	0	0	0	0.08	0	0	0.34	0	0
MIPGAN-I	Ensemble Features [20]	39.97	75.98	68.78	20.21	42.14	33.44	20.73	45.28	36.53
	Hybrid Features [41]	46.45	86.79	77.87	29.34	59.19	47.51	24.87	51.62	41.18
	Proposed	16.61	27.66	20.45	71.95	99.83	99.66	41.20	81.44	75.95
MIPGAN-II	Ensemble Features [20]	39.93	73.58	66.89	15.78	28.14	19.38	13.72	28.98	16.63
	Hybrid Features [41]	44.72	82.16	73.75	19.36	43.22	28.64	16.98	32.93	23.84
	Proposed	15.09	26.63	19.76	10.85	95.53	62.89	11.53	58.42	21.82

Appendix E. Unknown testing - MIPGAN-II trained

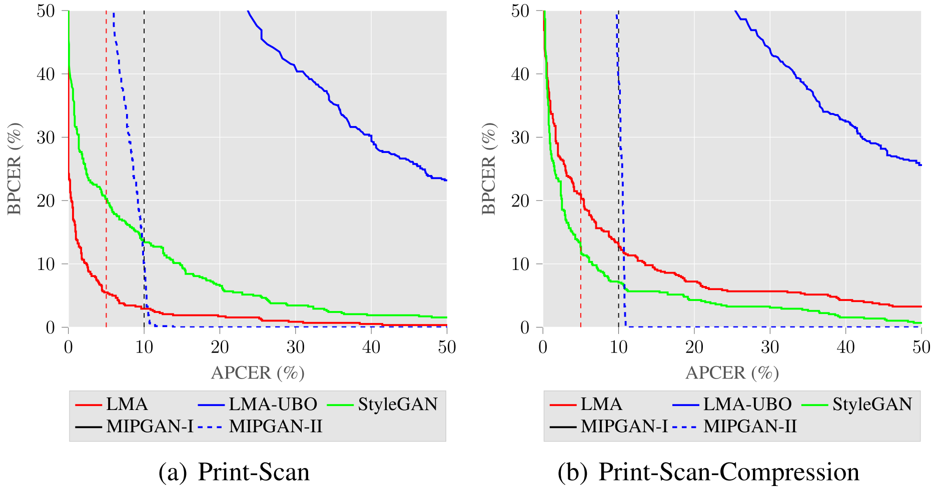
Similar to StyleGAN data, another set of experiments is conducted using MIPGAN-II as a training set, and the results are presented in Table E.5. We further note that the MIPGAN-II trained data does not generalize well on MIPGAN-I, StyleGAN or other landmark-based approaches when trained under the same settings as the rest of the others mentioned above (i.e., 30 epochs). We, therefore, conduct another set of experiments by increasing the epochs to 60 for both print-scan and print-scan-compression data. The results corresponding to this experiment are presented in Table E.6, and the DET curves for this set of experiments are illustrated in Fig. E.15. Finally, we note that the low performance of the MIPGAN-II trained model in Table E.5 reduces significantly when the number of epochs is increased, leading to better-generalized performance across the different unknown testing sets. However, one can note that the EER for the MIPGAN-II dataset itself does not increase marginally, but the BPCER for MIPGAN-II testing at APCER = 5% and APCER = 10% for print-scan and print-scan-compression increases. The observation can be noted from Fig. E.15 where a sudden increase can be noted around APCER = 10%, while the performance for both MIPGAN-I and MIPGAN-II decreases significantly with an increased number of epochs. However, the same model performs near-ideal at APCER = 20%, indicating the need for further investigations. A potential reason for this observation can be argued in the

Table E.5
Quantitative performance of MAD trained on MIPGAN-II data [9]. Results are noted in blue when proposed approach is superior over SOTA or equal to SOTA and noted in red when it is inferior to SOTA.

	Morphing	Digital			Print Scan			Print Scan Compression		
	Generation	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA	Ensemble Features [20]	13.08	29.15	15.78	4.28	3.94	2.22	4.28	3.61	2.22
	Hybrid Features [41]	40.14	77.7	67.23	5.49	5.48	2.4	7.21	10.98	4.15
	Proposed	50	82.47	74.57	11.22	17.53	12.37	20.27	38.32	30.58
LMA-UBO	Ensemble Features [20]	32.37	84.9	70.32	39.2	90.12	82.32	44.17	95.49	88.73
	Hybrid Features [41]	23.88	63.8	45.62	40.22	88.9	79.2	38.96	94.28	82.14
	Proposed	64.75	82.47	82.47	45.88	79.38	79.38	51.03	80.07	80.07
StyleGAN	Ensemble Features [20]	12.51	22.29	15.78	13.72	29.67	18.18	14.25	31.73	20.41
	Hybrid Features [41]	24.7	49.74	41.85	12.87	26.58	14.75	11.86	26.92	15.09
	Proposed	33.33	72.34	64.95	13.81	25.26	17.01	11.68	20.45	13.75
MIPGAN-I	Ensemble Features [20]	1.56	0.68	0.34	2.14	1.22	0.53	2.57	0.85	0.34
	Hybrid Features [41]	2.27	0.85	0.17	4.79	4.8	3.43	4.3	3.6	2.22
	Proposed	1.37	0.68	0.17	45.95	79.38	79.38	35.51	80.07	72.51
MIPGAN-II	Ensemble Features [20]	0	0	0	0	0	0	0	0	0
	Hybrid Features [41]	0	0	0	0	0	0	0	0	0
	Proposed	0.86	0.17	0.17	8.73	29.9	2.58	8.73	31.79	2.92

Table E.6
Quantitative performance of proposed MAD trained with MIPGAN-II [9] with increased number of epochs (60) for print-scan and print-scan-compression data.

Morphing	Print Scan			Print Scan Compression		
	EER	BPCER_20	BPCER_10	EER	BPCER_20	BPCER_10
LMA [43]	5.33	5.50	2.92	11.34	20.27	12.71
LMA-UBO [25]	35.17	67.01	67.01	36.04	66.15	66.15
StyleGAN [10]	12.51	20.27	13.40	8.16	11.68	7.04
MIPGAN-I [9]	53.13	67.01	67.01	59.46	66.15	66.15
MIPGAN-II [9]	10.02	57.39	9.28	10.70	66.15	38.49



(a) Print-Scan

(b) Print-Scan-Compression

Fig. E.15. DET curves obtained on proposed approach with MIPGAN-II data in training with 60 epochs for print-scan and print-scan-compression data.

light of over-fitting due to the limited number of samples. Therefore, alternative training strategies should be investigated to mitigate such behaviour.

References

[1] M. Ferrara, A. Franco, D. Maltoni, The magic passport, IEEE International Joint Conference on Biometrics, IEEE 2014, pp. 1–7.

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

[3] R. Raghavendra, K.B. Raja, S. Venkatesh, C. Busch, Transferable deep-cnn features for detecting digital and print-scanned morphed face images, 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE 2017, pp. 1822–1830.

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

[5] M. Gomez-Barrero, C. Rathgeb, U. Scherhag, C. Busch, Is your biometric system robust to morphing attacks? 2017 5th International Workshop on Biometrics and Forensics (IWBF), IEEE 2017, pp. 1–6.

[6] S. Venkatesh, R. Raghavendra, K. Raja, C. Busch, Face Morphing Attack Generation & Detection: A Comprehensive Survey, IEEE Trans. Technol. Soc. (2021).

[7] M. Ferrara, A. Franco, D. Maltoni, Face demorphing, IEEE Trans. Inf. Forensics Secur. 13 (4) (2018) 1008–1017.

[8] K. Raja, M. Ferrara, A. Franco, L.J. Spreeuwers, I. Batskos, F. de Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. Venkatesh, J.M. Singh, G. Li, L. Bergeron, S. Isadskiy, R. Ramachandra, C. Rathgeb, D. Frings, U. Seidel, F. Knopjes, R.N.J. Veldhuis, D. Maltoni, C. Busch, Morphing Attack Detection - Database, Evaluation Platform and Benchmarking, IEEE Trans. Inf. Forensics Secur. 16 (2020) 4336–4351.

[9] H. Zhang, S. Venkatesh, R. Raghavendra, K. Raja, N. Damer, C. Busch, MIPGAN-Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN, IEEE Trans. Biom. Behav. Identity Sci. (2021).

[10] S. Venkatesh, Z. Haoyu, R. Raghavendra, K. Raja, N. Damer, C. Busch, Can GAN Generated Morphs Threaten Face Recognition Systems Equally as Landmark Based Morphs? - Vulnerability and Detection, 2020 International Workshop on Biometrics and Forensics (IWBF), IEEE 2020, pp. 1–6.

[11] N. Damer, A.M. Saladié, A. Braun, A. Kuijper, MorGAN: recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network, 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE 2018, pp. 1–10.

[12] G. Borghi, A. Franco, G. Graffieti, D. Maltoni, Automated Artifact Retouching in Morphed Images With Attention Maps, IEEE Access 9 (2021) 136561–136579.

[13] NTNU, State of The Art Morphing Detection SOTAMD, 2019. URL:https://www.ntnu.edu/iik/sotamd.

[14] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, C. Busch, Face recognition systems under morphing attacks: A survey, IEEE Access 7 (2019) 23012–23026.

[15] N. Damer, S. Zienert, Y. Wainakh, A.M. Saladie, F. Kirchbuchner, A. Kuijper, A multi-detector solution towards an accurate and generalized detection of face morphing attacks, 22nd International Conference on Information Fusion, FUSION 2019, pp. 2–5.

[16] N. Damer, A.M. Saladié, S. Zienert, Y. Wainakh, P. Terhörst, F. Kirchbuchner, A. Kuijper, To detect or not to detect: the right faces to morph (2019).

[17] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhörst, A. Braun, A. Kuijper, Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts, German Conference on Pattern Recognition, Springer 2018, pp. 518–534.

[18] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, A. Uhl, Detection of Face Morphing Attacks Based on PRNU Analysis, IEEE Trans. Biom. Behav. Identity Sci. 1 (4) (2019) 302–317.

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

[20] S. Venkatesh, R. Raghavendra, K. Raja, C. Busch, Single Image Face Morphing Attack Detection Using Ensemble of Features, in: 23rd International Conference on Information Fusion, 2020, pp. 1–5.

[21] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, C. Busch, Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 280–289.

[22] U. Scherhag, C. Rathgeb, J. Merkle, C. Busch, Deep Face Representations for Differential Morphing Attack Detection, IEEE Trans. Inf. Forensics Secur. 15 (2020) 3625–3639.

[23] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, N.M. Nasrabadi, Attention aware wavelet-based detection of morphed face images, 2021 IEEE International Joint Conference on Biometrics (IJCB), IEEE 2021, pp. 1–8.

[24] S. Soleymani, A. Dabouei, F. Taherkhani, J. Dawson, N.M. Nasrabadi, Mutual Information Maximization on Disentangled Representations for Differential Morph Detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1731–1741.

[25] M. Ferrara, A. Franco, D. Maltoni, Face morphing detection in the presence of printing/scanning and heterogeneous image sources, IET Biom. 10 (3) (2021) 290–303.

[26] N. Damer, S. Zienert, Y. Wainakh, A.M. Saladié, F. Kirchbuchner, A. Kuijper, A Multi-detector Solution Towards an Accurate and Generalized Detection of Face Morphing

Attacks, 22th International Conference on Information Fusion, FUSION 2019, pp. 1–8.

[27] C. Seibold, A. Hilsmann, P. Eisert, Style Your Face Morph and Improve Your Face Morphing Attack Detector, in: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG), 2019, pp. 1–6.

[28] U. Scherhag, C. Rathgeb, C. Busch, Morph Detection from Single Face Image: A Multi-Algorithm Fusion Approach, Proceedings of the 2018 2nd International Conference on Biometric Engineering and Applications, ICBEA '18, Association for Computing Machinery, New York, NY, USA, ISBN: 9781450363945 2018, pp. 6–12, https://doi.org/10.1145/3230820.3230822 , URL: doi:10.1145/3230820.3230822.

[29] U. Scherhag, C. Rathgeb, C. Busch, Towards Detection of Morphed Face Images in Electronic Travel Documents, in: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), 2018, pp. 187–192.

[30] C. Seibold, W. Samek, A. Hilsmann, P. Eisert, Detection of face morphing attacks by deep learning, in: International Workshop on Digital Watermarking, 2017, pp. 107–120.

[31] A. Makrushin, C. Kraetzer, J. Dittmann, C. Seibold, A. Hilsmann, P. Eisert, Dempster-Shafer Theory for Fusing Face Morphing Detectors, in: 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5.

[32] C. Seibold, W. Samek, A. Hilsmann, P. Eisert, Accurate and robust neural networks for face morphing attack detection, J. Inf. Secur. Appl. 53 (2020), 102526.

[33] M. Ngan, P. Grother, K. Hanaoka, J. Kuo, Face Recognition Vendor Test (FRVT) Part 4: MORPH-Performance of Automated Face Morph Detection, Natl. Inst. Technol. (NIST), Tech. Rep. NISTIR 8292 (2021).

[34] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, N.M. Nasrabadi, Morph Detection Enhanced by Structured Group Sparsity, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 311–320.

[35] S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson, N.M. Nasrabadi, Differential morphed face detection using deep siamese networks, International Conference on Pattern Recognition, Springer 2021, pp. 560–572.

[36] C. Seibold, A. Hilsmann, P. Eisert, Feature Focus: Towards Explainable and Transparent Deep Face Morphing Attack Detectors, Computers 10 (9) (2021) 117.

[37] K. O'Haire, S. Soleymani, B. Chaudhary, P. Aghdaie, J. Dawson, N.M. Nasrabadi, Adversarially Perturbed Wavelet-based Morphed Face Generation, arXiv preprint arXiv:2111.01965 (2021).

[38] F. Peng, L.-B. Zhang, M. Long, Fd-gan: Face de-morphing generative adversarial network for restoring accomplice's facial image, IEEE Access 7 (2019) 75122–75131.

[39] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, A. Kuijper, PW-MAD: pixel-wise supervision for generalized face morphing attack detection, International Symposium on Visual Computing, Springer 2021, pp. 291–304.

[40] R. Raghavendra, S. Venkatesh, K. Raja, C. Busch, Detecting Face Morphing Attacks with Collaborative Representation of Steerable Features, in: IAPR International Conference on Computer Vision & Image Processing (CVIP-2018), 2018, pp. 1–7.

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

[42] boep-ubo, Best practice technical guidelines for Automated Border Control (ABC) systems, European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, 2015.

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

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

[45] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor, Asymmetric loss for multi-label classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 82–91.

[46] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737 (2017).

[47] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol. 1, IEEE 2005, pp. 947–954.

[48] International Civil Aviation Organization, Machine Readable Passports – Part 9 – Deployment of Biometric Identification and Electronic Storage of Data in eMRTDs, http://www.icao.int/publications/Documents/9303_p9_cons_en.pdf, 2015.

[49] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R.N. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, et al., Biometric systems under morphing attacks: assessment of morphing techniques and vulnerability reporting, 2017 International Conference of the Biometrics Special Interest Group (BIOSIG), IEEE 2017, pp. 1–7.

[50] DNP Imagingcomm America Corporation, DNP Printer, 2020. URL:http://dnpphoto.com/en-us/Products/Printers/DS820A.

[51] ISO/IEC JTC1 SC37 Biometrics, ISO/IEC 30107-3. Information Technology - Biometric presentation attack detection - Part 3: Testing and Reporting, International Organization for Standardization, 2017.

[52] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, E. Cabello, Border Control Morphing Attack Detection with a Convolutional Neural Network De-morphing Approach, IEEE Access (2020) 1–1.

[53] F. Peng, L.-B. Zhang, M. Long, FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplice's Facial Image, IEEE Access 7 (2019) 75122–75131, https://doi.org/10.1109/access.2019.2920713.

[54] M. Hildebrandt, T. Neubert, A. Makrushin, J. Dittmann, Benchmarking face morphing forgery detection: application of stirtrace for impact simulation of different processing steps, in: 2017 5th International Workshop on Biometrics and Forensics (IWBF), 2017, pp. 1–6.

[55] C. Seibold, A. Hilsmann, P. Eisert, Reflection analysis for face morphing attack detection, 2018 26th European Signal Processing Conference (EUSIPCO), IEEE 2018, pp. 1022–1026.

[56] C. Kraetzer, A. Makrushin, T. Neubert, M. Hildebrandt, J. Dittmann, Modeling Attacks on Photo-ID Documents and Applying Media Forensics for the Detection of Facial Morphing, in: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, 2017, pp. 21–32.

[57] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, B.D. Grad-CAM, Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 22–29.

[58] D. Saurabh, G. Harish, Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization, Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA 2020, pp. 1–5.

[59] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score-CAM: Score-weighted visual explanations for convolutional neural networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 24–25.

[60] Cognitec Systems GmbH, FaceVACS Technology - Version 9.4.2, 2020. URL:https://www.cognitec.com/facevacs-technology.html.

[61] Neurotechnology, Neurotech VeriLook SDK, 2020. URL:https://www.neurotechnology.com/verilook.html.

[62] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.