Conditional Identity Disentanglement for Differential Face Morph Detection

Sudipta Banerjee and Arun Ross
Michigan State University
banerj24@cse.msu.edu, rossarun@cse.msu.edu

Abstract

We present the task of differential face morph attack detection using a conditional generative network (cGAN). To determine whether a face image in an identification document, such as a passport, is morphed or not, we propose an algorithm that learns to implicitly disentangle identities from the morphed image conditioned on the trusted reference image using the cGAN. Furthermore, the proposed method can also recover some underlying information about the second subject used in generating the morph. We performed experiments on AMSL face morph, MorGAN, and EMorGAN datasets to demonstrate the effectiveness of the proposed method. We also conducted cross-dataset and cross-attack detection experiments. We obtained promising results of 3% BPCER @ 10% APCER on intra-dataset evaluation, which is comparable to existing methods; and 4.6% BPCER @ 10% APCER on cross-dataset evaluation, which outperforms state-of-the-art methods by at least 13.9%.

1. Introduction

Face morphing involves a continuous transition from the face image of one individual (source identity) to the face image of the second individual (target identity) [25]. The idea of morphing has been used for visual effects in entertainment videos. But recently, it has been demonstrated that face morphing can be used for adversarial purposes. Since a morphed face contains features from two individuals, it can successfully match both identities, thereby posing a security threat. See Figure 1. This problem is of practical concern due to two reasons: 1) the prolific use of biometric data in official documents for authentication in unattended border control systems (viz., face images in passports accepted at e-gates), and 2) the ease of access to face image editing software (e.g., FaceMorpher [2]). The idea of using morphed face image in identity documents was first postulated in [9].[1] Later, a real-world case of face morphing was reported where an art activist morphed her face image with that of the photograph of an EU Foreign Affairs Commissioner to apply for a German passport, prompting the authorities to reconsider using self-created digital photographs in identity documents [14]. See Figure 2. Face morph attack has piqued the interest of the research community and government agencies alike, leading to the investigation of methods that can not only produce realistic face morphs but also successfully detect such types of morphs. The EU-funded iMARS project is geared towards developing image morphing techniques and manipulation attack detection solutions for identification documents [3]. Note that current deepfake detectors cannot effectively discriminate between morphed and legitimate bonafide (non-morphed) images. Further, existing anti-spoofing solutions developed for presentation attacks are often not suited to detect morph attacks since the latter are *digital* alterations rather than *physical* presentations.

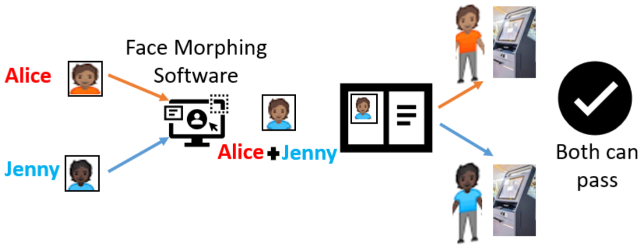


Figure 1: Illustration of face morphing exploited to provide access to two different subjects, Alice and Jenny. Both can use the same morphed image on the passport at an airport e-gate for access.

Face morph generation typically involves two kinds of approaches: 1) landmark-based approaches, and 2) generative model-based approaches. The first approach utilizes facial landmarks from two face images for aligning them using warping [12, 22]. Image blending principles are then used to combine the pixels in the overlapped regions to construct the morphed image. An optional post-processing step involving histogram equalization or Poisson blending may be applied to achieve visual realism and to remove ghosting

[1] Note that face morphing can be used in a positive manner also in privacy-preserving applications [17].

978-1-6654-3780-6/21/$31.00 ©2021 IEEE

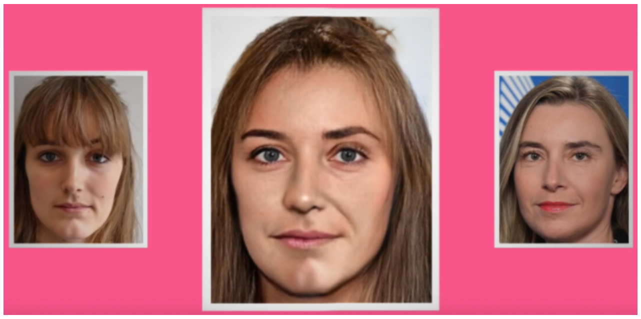


Figure 2: The morphed image in the middle that was used to apply for a passport in the EU was created using the photograph of an activist (left) and an official (right). Image courtesy: Peng!(MaskID) reproduced from [14].

artefacts. The second approach leverages the generative capability of adversarial networks to synthesize morphs. MorGAN [8], MIPGAN [28] and EMorGAN [7] are examples of such approaches. The final morphed image can then be inserted into an official document used at access control points. The morphs can be so realistic that visual inspection alone may not be sufficient for detecting them. See Figure 2.

Face morph attack detection (MAD) can be performed using two methods: 1) reference-free single image-based MAD, and 2) reference-based differential MAD. The former tries to address the following question. *Given a document face image can we determine whether it is morphed or not?* According to the NIST FRVT report [16], the best performing algorithm in this category results in 93.8% APCER @ 10% BPCER on high quality morphs in "Dataset:Manual" [16]. Here, bonafide presentation classification error rate (BPCER) denotes the proportion of bonafide (non-morphed) images incorrectly classified as morphed images, and attack presentation classification error rate (APCER) denotes the proportion of morphed images incorrectly classified as non-morphed images. The second method tries to address the following question. *Given a document face image, and a live capture as a reference image, can we determine whether the document image is morphed or not?* According to the NIST FRVT report [16], the best performing algorithm in this category results in 9.1% APCER @ 10% BPCER on high quality morphs in "Dataset:Manual". It is evident, based on these results, that the state-of-the-art performance requires significant improvement. The readers are referred to surveys in [4, 27] for a comprehensive overview of face morph generation and detection. In this paper, we focus on developing a novel *differential MAD method* to advance the state of the art. **The novelty of the method lies in formulating the differential MAD problem using an information theoretic framework for a sound detection strategy.**

The remainder of the paper is organized as follows. Section 2 describes existing differential MAD strategies. Section 3 describes the proposed method. Section 4 outlines the experiments. Section 5 reports and analyzes the results. Finally, section 6 concludes the paper with summary and future directions.

2. Related Work

Differential MAD strategies use a reference-based approach. Here, the trusted live face image of a subject taken during the time of acquisition (called the reference image) is used along with the document face image (in the passport, for example) to determine whether the latter is morphed or non-morphed. The first known work on differential MAD performs "demorphing" which uses the difference computed between the document image and the reference image to ascertain whether the document image is morphed or not, and alerts the officer for additional inspection, if necessary [10]. Demorphing also tries to uncover the identity of the second subject. Later, GAN-based de-morphing has been proposed in the literature [18]. Another class of techniques utilizes features such as BSIF or the disparity between landmarks, along with a classifier, to detect morphs differentially [20, 6, 19]. The detection performance is further enhanced by the use of deep face representations [21]. Recently, a method that uses appearance and landmark disentanglement modules for differential MAD was proposed in [24]. It uses the disentanglement module to learn complementary information, and contrastive loss to boost detection performance. An approach referred to as Focused Layer-Wise Relevance Propagation (FLRP) aims at explaining the decision made by the deep neural network in detecting morphs for better interpretability [5].

3. Proposed Method

3.1. Conditional GAN

The proposed method requires a tool that can convert a source image, $\boldsymbol{X}$, which is the 'document' image, to a target image, $\boldsymbol{Y}$, which is the 'reference' image. This can be achieved via image translation guided by a conditional generative adversarial network (cGAN) [13]. It uses the following objective function.

$$\begin{aligned} G^* = \arg\min_G \max_D \{ & \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y})}[\log D(\boldsymbol{X},\boldsymbol{Y})] \\ & + \mathbb{E}_{(\boldsymbol{X},\boldsymbol{z})}[\log(1 - D(\boldsymbol{X}, G(\boldsymbol{X},\boldsymbol{z})))]\} \\ & + \lambda \mathbb{E}_{(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{z})}[\|(\boldsymbol{Y} - G(\boldsymbol{X},\boldsymbol{z}))\|_1] \end{aligned} \tag{1}$$

Here, D refers to the discriminator and G refers to the generator. The two inputs to the network are $\boldsymbol{X}$ (document image) and $\boldsymbol{Y}$ (reference image). The reference image is always assumed to be a bonafide since it is captured "live". Now the network has to learn to translate

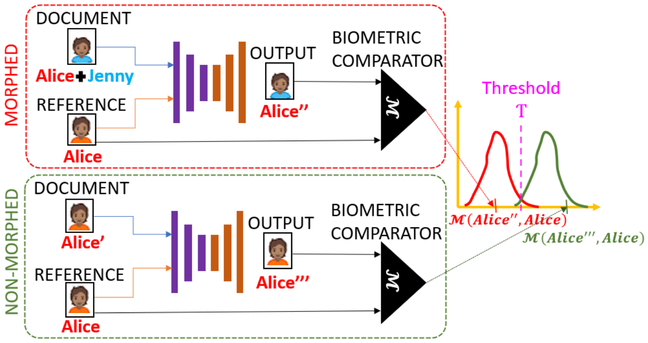


Figure 3: The proposed method uses a conditional disentanglement framework to discriminate between morphed and non-morphed document images conditioned on the reference image. The biometric comparator scores can be compared against a suitable threshold, $\mathcal{T}$, for differential morph attack detection.

from $\boldsymbol{X}$ to $\boldsymbol{Y}$. The Gaussian random noise vector $\boldsymbol{z}$ regularizes the training and diversifies the output. If $\boldsymbol{X}$ is non-morphed, then the task of the generator is to simply reconstruct itself with some additional variations due to age, pose, illumination and expression, which can be deduced from $\boldsymbol{Y}$. However, if $\boldsymbol{X}$ is morphed, the challenge is two-fold: (i) remove traces of the second identity, and (ii) incorporate variations related to pose, age, expression and appearance. We train the network with both types of examples: $(\boldsymbol{X}_{non-morphed}, \boldsymbol{Y})$ and $(\boldsymbol{X}_{morphed}, \boldsymbol{Y})$, i.e., (non-morphed source and bonafide target images) and (morphed source and bonafide target images). **However, the method does not require any labels (supervision) regarding which pairs correspond to morphed source images and which pairs correspond to non-morphed source images.** The idea is to make the network automatically learn to disentangle the composite (morphed) image conditioned on the reference image. If the image is not a composite, the task is simpler, it has to generate a variant of the source image.

3.2. Rationale of the Proposed Method

The majority of the literature poses differential MAD as a supervised classification problem, with the underlying premise that discriminative features can be deduced using either pre-defined filters (hand-crafted) or automatically learned filters (deep-learning). In contrast, we make the case that the task of differential MAD can be framed in an information theoretic framework that describes the translation from the document image to the reference image.

$$\mathcal{H}(\boldsymbol{X} \mid \boldsymbol{Y}) = \mathcal{H}(\boldsymbol{Y}) - \mathcal{H}(\boldsymbol{X}, \boldsymbol{Y}) \tag{2}$$

If $\boldsymbol{X}$ is non-morphed, then $\boldsymbol{Y} = [\boldsymbol{X} + \delta]$, where δ captures the intra-class variations between document and reference images belonging to the same subject. Therefore, the conditional entropy can be formulated as,

$$\mathcal{H}(\boldsymbol{X} \mid [\boldsymbol{X} + \delta]) = \mathcal{H}([\boldsymbol{X} + \delta]) - \mathcal{H}(\boldsymbol{X}, [\boldsymbol{X} + \delta]) \tag{3}$$

Clearly the uncertainty associated with inferring $\boldsymbol{X}$ from a slightly noisy or altered version of itself (δ can be interpreted as additive noise) will be low. On the contrary, for a morphed image comprising two statistically independent distinct subjects, the uncertainty will be relatively higher due to the presence of a second identity.

$$\begin{aligned}\mathcal{H}([\boldsymbol{X}_1, \boldsymbol{X}_2] \mid [\boldsymbol{X}_1 + \delta]) = \\ \mathcal{H}([\boldsymbol{X}_1 + \delta]) - \mathcal{H}([\boldsymbol{X}_\mathbf{1}, \boldsymbol{X}_\mathbf{2}], [\boldsymbol{X}_1 + \delta])\end{aligned} \tag{4}$$

In Eqn. (4), we assume subject $\boldsymbol{X}_1$ appears at the verification checkpoint, so the uncertainty arises due to $\boldsymbol{X}_2$. A similar situation arises when the roles are reversed. Therefore, we expect $\mathcal{H}(\boldsymbol{X}_{morphed} \mid \boldsymbol{Y}) > \mathcal{H}(\boldsymbol{X}_{non-morphed} \mid \boldsymbol{Y})$. Here, we interpret entropy loosely as the disparity between the output of the cGAN and the reference image. Next, we describe the two steps in which we implement the proposed method.

In the first step, we use the cGAN to translate the source image (document) to the target image (reference): $\boldsymbol{X} \rightarrow \boldsymbol{Y}$. If $\boldsymbol{X}$ is non-morphed and represents the same individual as $\boldsymbol{Y}$, then the output (translated) image will be *more similar* to $\boldsymbol{Y}$. On the other hand, if $\boldsymbol{X}$ is morphed, implying that it comprises two identities $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, then the output image will be *less similar* to $\boldsymbol{Y}$. The translation will force the network to implicitly learn to disentangle identities. This is because, in order to translate from the source image (two identities if morphed) to the target image (only one identity), i.e., $[\boldsymbol{X}_1, \boldsymbol{X}_2] \rightarrow \boldsymbol{X}_2$ (or $\boldsymbol{X}_1$), it will try to remove traces pertaining to the second identity not present in the reference image, i.e., $\boldsymbol{X}_1$ (or $\boldsymbol{X}_2$), thereby striving to decouple the two identities present in the morphed image. We refer to this as *conditional identity disentanglement*, where we disentangle the identities from a morphed image conditioned on a reference image. A fortuitous outcome of this process is that we can use the same method for deciphering information about the second subject from the morphed image. This is the novelty of the proposed method. By positing the differential morph detection problem using an information theoretic framework, not only can we detect morphs, but also disentangle the identities. But how do we quantify the disparities between the output of the cGAN and the reference image to deduce whether a document image is morphed or not?

In the second step, we use the output, $\boldsymbol{O} = G(\boldsymbol{X}, \boldsymbol{z})$, of the cGAN, where $\boldsymbol{z}$ is the random noise vector, and compare it with the reference image, $\boldsymbol{Y}$, using a biometric comparator, $\mathcal{M}$. The score produced after the comparison can then be compared against a user-defined threshold, $\mathcal{T}$, that regulates error rates for the intended application to make the final decision:

$$\mathcal{M}(G(\boldsymbol{X}, \boldsymbol{z}), \boldsymbol{Y}) = \mathcal{M}(\boldsymbol{O}, \boldsymbol{Y}); \tag{5}$$

$$\boldsymbol{X} = \begin{cases} \text{Morphed}, & \text{if } \mathcal{M}(\boldsymbol{O}, \boldsymbol{Y}) < \mathcal{T}, \\ \text{Non-morphed}, & \text{otherwise.} \end{cases} \tag{6}$$

Figure 3 outlines the proposed method.

4. Experiments

We describe the (i) datasets, (ii) conditional GAN setup, and (iii) experimental protocols used in this work.

4.1. Dataset

We used three datasets in this work. (i) **AMSL face morph dataset** [15, 1]: It contains images from 102 subjects captured with neutral as well as smiling expressions. There are 2,175 morphed images corresponding to 92 subjects created using a landmark-based approach. In our problem formulation, the document image can be either (a) an image with neutral expression which will be considered as the bonafide, or (b) a morphed image. The reference image is a trusted live capture (bonafide) corresponding to the image of the same subject but with a smiling expression. (ii) **MorGAN dataset** [8]: It contains 500 bonafide images. Two morphs are generated using a generative network for each of these bonafide images from the two subjects most similar to the bonafide image resulting in 1,000 morphed images, which were split into train and test sets. (iii) **EMorGAN dataset** [7]: It uses a cascaded image enhancement network to improve the quality of the morphed images synthesized using MorGAN, and has the same train and test split as [8].

4.2. Implementation Details

We explored different options for conditional GANs (cGANs) in the literature and found PIX2PIX [11] to be a suitable choice for this work. PIX2PIX uses a cGAN to translate images from one domain (e.g., sketch) to another domain (e.g., photorealistic images). PIX2PIX [11] follows $Conv \rightarrow BatchNorm \rightarrow ReLU$ architecture both in the generator and discriminator. The discriminator loss function minimizes the difference between the real and the fake images. The generator loss function maximizes the log-likelihood of the generated images while ensuring they are as close as possible to the target images using $\mathcal{L}_1$ loss. Readers are referred to [11] for additional details about the implementation. We used mini-batch stochastic gradient descent optimization algorithm with Adam solver at an initial learning rate of 2×10^{-4} and momentum parameter of 0.5, and trained for 50/100/200 epochs (compared to 600K iterations in [24]).

4.3. Experiments

We conducted three experiments designed to answer the following research questions.

Experiment 1: *How does the proposed differential MAD method perform?*
To answer the above question, we trained on the AMSL dataset using 120 pairs of document and reference images out of which 60 pairs were morphed document images and the other 60 pairs were bonafide document images. These 60 subjects constitute 65% of the 92 subjects, while the test set comprised of 778 images (745 morphed and 33 bonafide images) corresponding to the remaining 35% of the subjects. **During training, no label is required to indicate which pairs correspond to morphed images and which pairs correspond to non-morphed images.** We feed the images generated by our network and the target reference images to a face recognition system (a COTS face comparator) and use the scores to determine whether a document image is morphed or not. We selected a subset of training images provided by the authors in the MorGAN [8] and the EMorGAN [7] datasets. We trained on 120 pairs (60 pairs of bonafide document images and 60 pairs of morphed document images), and used the entire test set for evaluation in both cases.

Experiment 2: *Is the proposed method generalizable under cross-dataset and cross-attack scenarios?*
To answer the above question, we trained on one dataset (one type of attack) and tested on the remaining two datasets (other types of attacks). AMSL dataset represents a landmark-based morph attack while MorGAN and EMorGAN datasets represent generative model-based attacks.

Experiment 3: *Can the method be used for recovering some discriminative information about the second subject?*
To answer the above question, we used a different training strategy compared to morph detection. In this experiment, we used the pixel-wise difference image computed between the document and the reference image from the AMSL dataset as the source image, and the document image as the target image. The intuition behind this strategy is to ensure that the network learns to map the residual (reminiscent of the second subject contributing to the morph) to the morphed target image.

5. Results and Analysis

We report both qualitative and quantitative results of the proposed method. We report the results in terms of the metrics predominantly used in the morph detection literature: BPCER @ APCER of 10% following [24] as well as APCER @ BPCER of 10% following [16].

Results from Experiment 1: Table 1 reports the performance of the proposed method. We would like to reiterate that the proposed method was not trained to discrim-

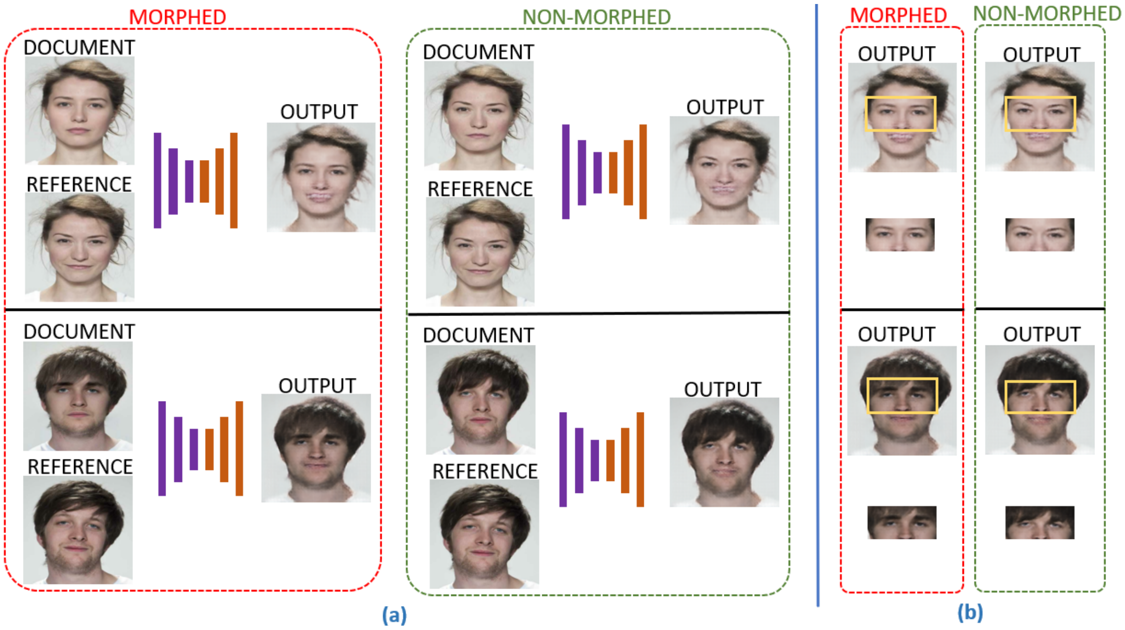


Figure 4: (a) Illustration of conditionally disentangled outputs generated using the proposed method for two subjects (top and bottom) belonging to the AMSL dataset when their document images were morphed (red) and non-morphed (green). (b) Zoomed-in view of the periocular regions of the generated outputs to emphasize the difference between morphed and non-morphed images.

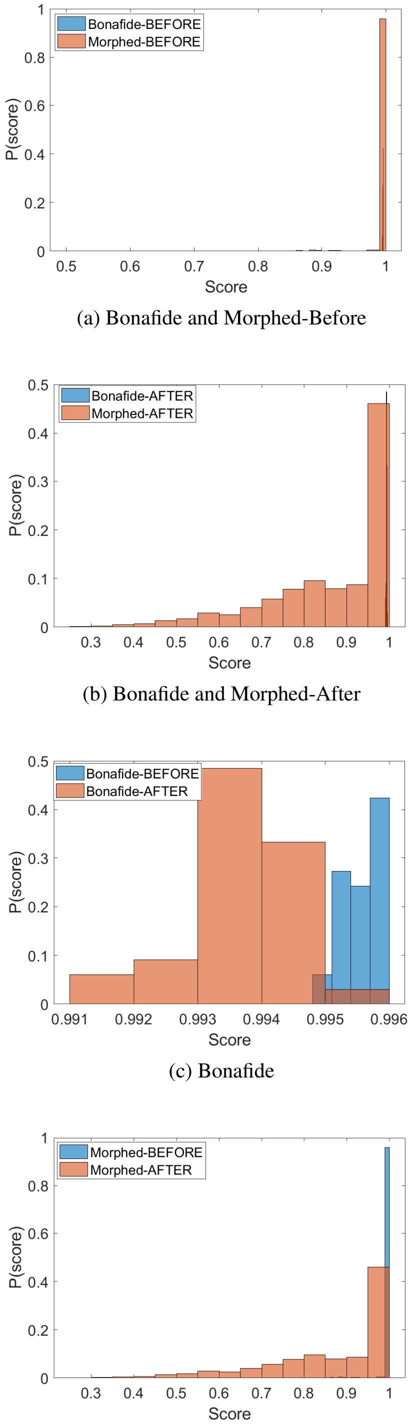
inate between non-morphed and morphed images, i.e., during training the method does not require labels corresponding to morphed and non-morphed image pairs. In spite of that, the proposed method achieves comparable results to some extent with state-of-the-art methods that require explicit supervision (see Table 2). Qualitatively, we present the results in Figure 4. In order to demonstrate that the proposed method is actually disentangling subject identities, we present the score distributions produced by the COTS comparator, when comparing test images belonging to the AMSL dataset (see Figure 5). In Figure 5(a), the distributions corresponding to morphed and bonafide images completely overlap (BEFORE disentanglement). In Figure 5(b), the distribution corresponding to morphed images move towards zero (AFTER disentanglement). Here, the scores are similarity scores normalized to $[0, 1]$. To illustrate how disentanglement has a pronounced effect on morphed images compared to bonafide images, we present the score distributions of the *bonafide* images before and after disentanglement in Figure 5(c), and that of the *morphed* images before and after disentanglement in Figure 5(d). The bonafide scores are closely located (note the range of the scores in the x-axis in Figure 5(c)). In contrast, if the document image is morphed, the network tries to disentangle the identity traces belonging to the second subject which causes a significant shift in the scores (see Figure 5(d)). We further computed the interquartile range (IQR) as a 'measure of dispersion' between the bonafide and morphed score distributions. This was done for bonafide and morphed scores both before and after disentanglement. Then we computed the ratio of their IQRs and observed that the ratio for morphed scores was ∼6 times higher than that of the bonafide scores. $\left(\frac{IQR_{morphed-AFTER}}{IQR_{morphed-BEFORE}} > 5.9 \times \frac{IQR_{bonafide-AFTER}}{IQR_{bonafide-BEFORE}}\right)$.

Table 1: BPCER(%) @ APCER=10% (left of the forward slash) and APCER(%) @ BPCER=10% (right of the forward slash).

Train \ Test	AMSL	MorGAN	EMorGAN
AMSL (50 epochs)	6.1/5.2	4.6/2.0	4.4/0.8
MorGAN (100 epochs)	63.6/60.3	8.6/5.8	9.4/9.3
EMorGAN (200 epochs)	25.8/47.7	28.1/38.4	29.7/39.4

We observed from the experiments that the AMSL dataset required the least number of epochs followed by MorGAN and EMorGAN. In the case of AMSL, the BPCER improved from 6.1% to 3% upon increasing the number of training epochs from 50 to 200. But the BPCER remained the same when tested on the MorGAN and EMorGAN datasets.

Results from Experiment 2: Table 2 compares the performance of the proposed method with the two baselines [24, 21]. The proposed method achieves



(a) Bonafide and Morphed-Before

(b) Bonafide and Morphed-After

(c) Bonafide

(d) Morphed

Figure 5: Variations in score distributions before and after conditional disentanglement in the AMSL dataset indicating successful disentanglement in morphed document images (Experiment 1).

Table 2: BPCER(%) @ APCER=10%. The best performing results for the proposed method are compared with the baseline results indicated within parentheses ([24], [21]). The baseline results are taken from [24].

Train \ Test	AMSL	MorGAN
AMSL	3.0 (**2.2**, 3.3)	**4.6** (18.5, 24.5)
MorGAN	63.6 (**8.8**, 14.9)	**8.5** (**8.5**, 12.4)

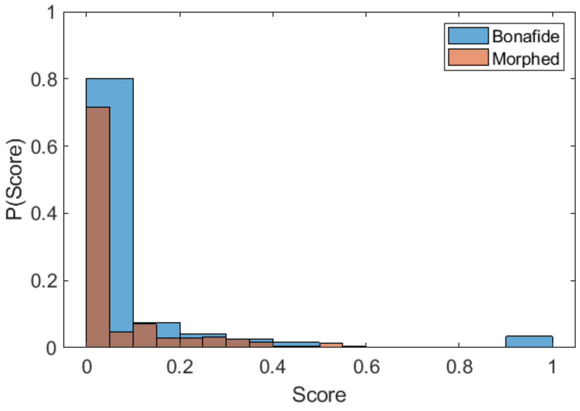


Figure 6: The COTS face comparator scores corresponding to bonafide and morphed test images belonging to the MorGAN dataset (Experiment 2).

BPCER=4.6% @ APCER=10%, which outperforms existing methods ([24]: BPCER=18.5% @ APCER=10% and [21]: BPCER=24.5% @ APCER=10%) by a considerable margin when trained on AMSL but tested on MorGAN dataset. However, training on the MorGAN dataset results in poor performance when tested on AMSL dataset. We investigated this issue further and observed an interesting phenomenon. We computed the Pearson correlation coefficient between the document and reference images in the AMSL test set, and when averaged across all the images, it resulted in a mean and standard deviation of 0.91 ± 0.03, while for the MorGAN test set, the mean and standard deviation were 0.48 ± 0.24, which is almost half of that on the AMSL dataset. We also computed the biometric utility of the MorGAN test set images using the COTS comparator and observed that the biometric similarity scores are low (see Figure 6) for both bonafide and morphed images. We suspect the the low degree of correlation between the document and reference images, and the overall low biometric matching utility of the images in the MorGAN dataset, may be responsible for the lower performance of the proposed method in this case. Our method requires adequately high degree of correlation between the document and reference images for successful identity disentanglement.

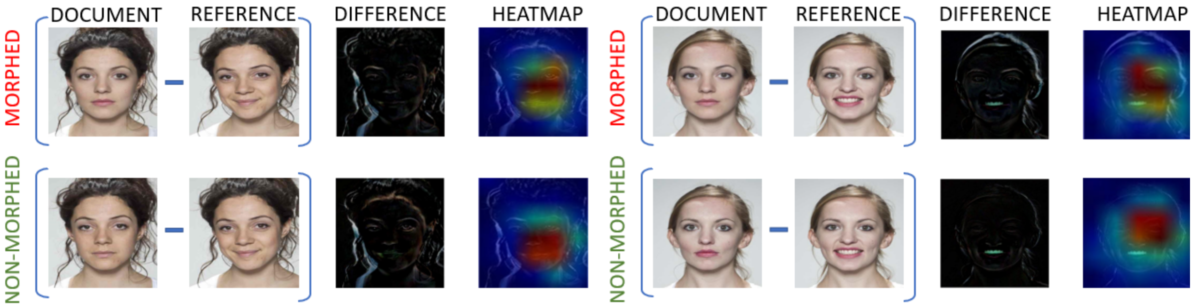


Figure 7: Heatmap visualization of the "difference images" for morphed and non-morphed inputs corresponding to two subjects. Note that the difference images contain some discernible information that can be harnessed for recovering some unique information about the second constituent in the morphed image (Experiment 3).

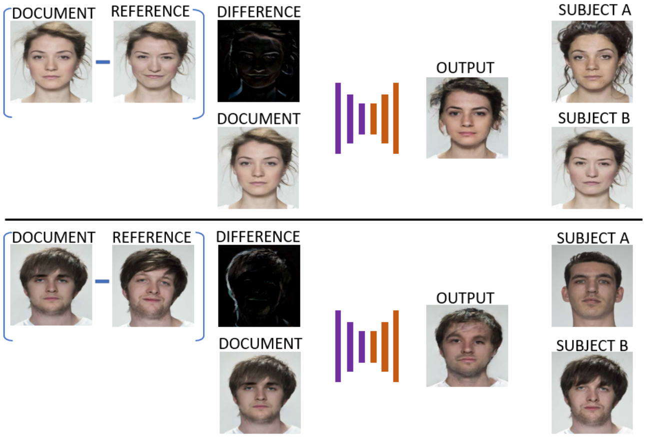


Figure 8: Illustration of outputs generated using the proposed method for decoding information about the "other" subject used in morph generation (Experiment 3). Note in both cases the output is visually more similar to Subject A (second subject) compared to Subject B (anchor subject present in reference image).

Results from Experiment 3: In order to recover some information about the second subject who contributed to the morphed image, we utilized the pixel-wise difference images. However, we first needed to ascertain whether the difference images contained any useful information at all. Therefore, we used a similarity visualization technique [26] to visualize whether the difference image retained any useful information. Figure 7 depicts the visualization in terms of heatmaps (saliency maps) which shows that the difference images contain discriminative information which can be harnessed for decoding the second subject. Figure 8 illustrates examples of images generated using the proposed method that *qualitatively* appear similar to the second subject (Subject A) compared to the anchor subject (Subject B); the anchor subject pertains to the identity in the reference image. Further, we also analyzed the similarity between the generated outputs and the constituent subjects *quantitatively* using a COTS face comparator. The biometric similarity between the generated outputs and the second subject, as assessed using the true match rate at a false match rate of 1%, was 23.4% higher than that between the generated outputs and the first (anchor) subject. Therefore, we note that the proposed method displays tremendous promise for decoding the second identity.

6. Summary and Future Work

In this paper, we proposed a novel differential face morph attack detection framework that disentangles subject identities from a morphed document image conditioned on the trusted reference image. In contrast to existing classification-based approaches, the proposed method formulates the differential MAD problem using an information theoretic framework. We used a conditional generative network to produce an output image from the input (bonafide or morphed) document image. Next, we compared the output image with the reference image using a biometric comparator. The ensuing score is then used for detecting morphs. The method demonstrates a promising first step across different kinds of morph attacks without requiring any supervision about the type of morphing. We achieved best performing results of 3% BPCER @10% APCER on intra-dataset evaluation and 4.6% BPCER @ 10% APCER on cross-dataset evaluation (Table 2). We observed that the proposed method needs well-correlated document and reference image pairs acquired in controlled settings for its robust operation (Section 5). This is a reasonable requirement given that the success of face morphing is enhanced under these conditions. Furthermore, we used the proposed method to recover some characteristics about the second subject (different from the first subject whose reference image was used). Future work will involve improving the task of de-morphing as well as studying the applicability of the proposed technique to other modalities [23].

References

[1] AMSL Face Morph Image Data Set. https://omen.cs.uni-magdeburg.de/disclaimer/index.php. [Online accessed: 15th February, 2021]. 4

[2] Face Morpher. http://www.facemorpher.com/. [Online accessed: 1st July, 2021]. 1

[3] image Manipulation Attack Resolving Solutions (iMARS). https://cordis.europa.eu/project/id/883356. [Online accessed: 2nd April, 2021]. 1

[4] K. B. Raja et al. Morphing Attack Detection - Database, Evaluation Platform and Benchmarking. *IEEE Transactions on Information Forensics and Security*, 2020. 2

[5] P. E. Clemens Seibold, Anna Hilsmann. Focused LRP: Explainable AI for Face Morphing Attack Detection. *IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 88–96, 2021. 2

[6] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhörst, A. Braun, and A. Kuijper. Detecting Face Morphing Attacks by Analyzing the Directed Distances of Facial Landmarks Shifts. *German Conference on Pattern Recognition*, 2018. 2

[7] N. Damer, F. Boutros, A. M. Saladie, F. Kirchbuchner, and A. Kuijper. Realistic Dreams: Cascaded Enhancement of GAN-generated Images with an Example in Face Morphing Attacks. *IEEE 10th International Conference on Biometrics Theory, Applications and Systems*, pages 1–10, 2019. 2, 4

[8] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper. MorGAN: Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network. *IEEE 9th International Conference on Biometrics Theory, Applications and Systems*, pages 1–10, 2018. 2, 4

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

[10] M. Ferrara, A. Franco, and D. Maltoni. Face Demorphing. *IEEE Transactions on Information Forensics and Security*, 13:1008–1017, 04 2018. 2

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017. 4

[12] A. Makrushin, T. Neubert, and J. Dittmann. Automatic Generation and Detection of Visually Faultless Facial Morphs. In *VISIGRAPP*, 2017. 1

[13] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *ArXiv*, abs/1411.1784, 2014. 2

[14] M. Monroy. Laws against morphing. https://digit.site36.net/2020/01/10/laws-against-morphing/. Appeared in Security Architectures and the Police Collaboration in the EU 10/01/2020 [Online accessed: 1st April, 2021]. 1, 2

[15] T. Neubert, A. Makrushin, M. Hildebrandt, C. Krätzer, and J. Dittmann. Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images. *IET Biometrics*, 7:325–332, 2018. 4

[16] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo. Face Recognition Vendor Test (FRVT) Part 4: MORPH - Performance of Automated Face Morph Detection. *NISTIR 8292 Draft Supplement*, April 16, 2021. 2, 4

[17] A. A. Othman and A. Ross. Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity. In *European Conference on Computer Vision Workshops*, 2014. 1

[18] F. Peng, L. B. Zhang, and M. Long. FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplice's Facial Image. *IEEE Access*, 7:75122–75131, 2019. 2

[19] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch. Detecting Morphed Face Images Using Facial Landmarks. In *International Conference on Image and Signal Processing*, 2018. 2

[20] U. Scherhag, C. Rathgeb, and C. Busch. Towards Detection of Morphed Face Images in Electronic Travel Documents. In *IAPR 13th International Workshop on Document Analysis Systems*, pages 187–192, 2018. 2

[21] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch. Deep Face Representations for Differential Morphing Attack Detection. *IEEE Transactions on Information Forensics and Security*, 15:3625–3639, 2020. 2, 5, 6

[22] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert. Detection of face morphing attacks by deep learning. In C. Kraetzer, Y.-Q. Shi, J. Dittmann, and H. J. Kim, editors, *Digital Forensics and Watermarking*, pages 107–120, 2017. 1

[23] R. Sharma and A. Ross. Image-level Iris Morph Attack. *IEEE 28th International Conference on Image Processing*, 2021. 7

[24] S. Soleymani, A. Dabouei, F. Taherkhani, J. Dawson, and N. M. Nasrabadi. Mutual Information Maximization on Disentangled Representations for Differential Morph Detection. *IEEE Winter Conference on Applications of Computer Vision*, pages 1731–1741, 2021. 2, 4, 5, 6

[25] M. Steyvers. Morphing techniques for manipulating face images. *Behavior Research Methods, Instruments, & Computersc*, 31:359–369, 06 1999. 1

[26] A. Stylianou, R. Souvenir, and R. Pless. Visualizing Deep Similarity Networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 2029–2037, 2019. 7

[27] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch. Face Morphing Attack Generation & Detection: A Comprehensive Survey. *IEEE Transactions on Technology and Society*, 2021. 2

[28] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch. MIPGAN - Generating Robust and High Quality Morph Attacks Using Identity Prior Driven GAN. *IEEE Transactions on Biometrics*, 2021. 2