Patch-Based Morphing Attack Detection

YULING JIANG
Computer Science, University of Twente
Supervised by Luuk Spreeuwers

***Abstract*—The morphing attack poses a significant threat to face recognition systems, as it undermines the unique link between identity and identification documents. Therefore, the need for morphing attack detection is imperative. In this paper, we proposed a novel patch-based morphing attacks detection approach. The facial regions from the images were cropped and divided into 30 patches. This methodology facilitates a straightforward expansion of the dataset size. We conducted a comprehensive analysis comparing different combinations of feature extraction networks and score fusion mechanisms. The findings demonstrate that the utilization of Se_Resnet50 as the feature extractor, combined with either the average or machine learning score fusion method, produces satisfactory results during the test phase. In particular, the D-EER for intra-dataset tests is 0%, and the highest D-EER observed for cross-dataset tests is merely 12.1%. However, when conducting cross-dataset testing, morphs generated with STYLEGAN2 exhibit an exception to this trend. Subsequently, extensive experiments with the optimal combination were conducted to investigate the influence of various training settings on the outcomes. The findings unveiled that when subjected to images generated by STYLEGAN2 from distinct datasets, a model exclusively trained on STYLEGAN2-generated images exhibited enhanced capabilities in generalization. Furthermore, there are certain similarities in the artifacts observed in both landmark-based and GAN-based morphed images.**

I. Introduction

Due to its non-intrusive nature, the Face Recognition System (FRS) finds extensive application across diverse scenarios, including surveillance, border control, and access control. Automated border control (ABC) deployed at airports is one of the most critical applications of face biometrics, where identity verification will be carried out by comparing an image stored in an electronic Machine Readable Travel Document (eMRTD) such as an e-passport with a live captured image.

However, FRSs have the inability to accurately detect manipulated images, especially for those made by morphing technology, which has now been demonstrated to be a potential threat to face verification scenarios. Multiple studies [1]–[3] have demonstrated the vulnerability of FRSs to morphing attacks, and the results are not encouraging. Face morphing manipulation allows for the combination of facial features of two subjects into one image. This enables the potential for two individuals who bear resemblance to each other to share a single identity document, thereby disrupting the exclusive association between an individual and their corresponding identity document. Fig. 1 shows an example of the resulting morph. Subject 1 and subject 2 generate the morphed image, which visually resembles both two contributing subjects.

In a recent investigation [1], an initial study uncovered the possibility of an attack carried out through the utilization of a morphed image within the ABC context. This morphing-based attack demonstrates practical feasibility, as certain regions permit citizens to submit ID photographs for the acquisition of official identification documents, thus presenting a substantial opportunity for such attacks. Moreover, existing research has found that it is inherently challenging for humans to recognize unfamiliar faces in small-sized pictures [4]. Faced with morphed images, even observers with prior knowledge are unable to achieve successful verification [5]. Although live enrollment would be an ideal solution to prevent forgery, it is not available in all countries. Additionally, issued electronic documents that contain biometric features also have potential security risks [4].

Fig. 1: Example of face morphing

In spite of the high severity, the process of generating morphed images does not require any specialized knowledge. There are a plethora of free tools available on the internet that allow a criminal to generate high-quality falsified images, such as FaceMorpher[1] and FaceFusion[2].

Therefore, under such circumstances, the adoption of a morph attack detection system becomes indispensable. Traditionally, most existing morph attack detection systems have relied on processing complete facial images to make decisions. In this study, we propose and investigate a groundbreaking method for morph detection that is based on patch segmentation. Instead of directly feeding the entire image into the detection system for evaluation, our approach assigns scores to small square regions extracted from the face. By segmenting the complete image, we can substantially increase the dataset size. We compare three feature extractors, namely Resnet50, Se_Resnet50, and VGG19, along with three score combination methods (average, majority vote, machine learning) in the patch-based detection scenario. The results of the intra-dataset test show excellent performance when utilizing the network architectures Se_Resnet50 and Resnet50. Furthermore, the cross-dataset test results based on Se_Resnet50 and VGG19 are also promising, except for morphs generated with STYLEGAN2. No matter which feature extractor is used, if the trained

[1] https://github.com/alyssaq/face_morpher
[2] www.wearemoment.com/FaceFusion

model proves effective on the test set, implementing a fusion strategy can obviously improve detection accuracy. We also conduct extensive experiments with the optimal combination to assess the performance disparities under different training settings. In summary, we investigate one overall question and a few sub-questions:

Research Question 1: How to employ patch-based methodologies in DMAD scenario?

Research Question 2: Which combination of network architecture and score fusion method is best suited for patch-based morph detection?

Research Question 3: What disparities manifest in the detection performance of models generated through distinct training settings?

The paper is organized as follows. In Section II, the process of face morphing and existing detection methods are discussed. Additionally, an overview of patch-based approaches in face-related areas is provided. Section III introduces a novel patch-based detection scheme. The subsequent Section IV provides a detailed exposition of our experiment's design. This section further presents an analysis of the obtained results. The last section V closes the paper with the conclusion and future work.

II. Related Work

A. Face Morphing Generation

Numerous studies have been conducted to explore and develop techniques for generating morph images. Broadly, it can be categorized into two directions: landmark-based morph image generation and GAN-based morph image generation.

In the context of landmark-based methods, the initial step involves determining the coordinates of prominent facial components in two images (I_0, I_1), such as the mouth, eyes, and nose. Manual annotation is the most direct and accurate approach, but it can be time-consuming. Alternatively, automatic annotation algorithms can be employed, although they may not provide the same level of precision. A commonly utilized landmark detector is Dlib [6]. Once the landmarks are determined, geometric warping is performed, often employing Delaunay triangulation [7]. During the warping process, the contribution of I_0 and I_1 is controlled by a parameter known as α_w. Finally, texture blending is applied, where another parameter α_b governs the contribution of I_0 and I_1 during the blending stage. This step combines texture details from the original images, resulting in a visually appealing and seamless morph image.

To overcome the issue of inaccurate landmark annotations, researchers have proposed a deep learning-based approach for morph generation. Unlike the landmark-based method, which necessitates the manual annotation of reference points, the GAN-based method employs a projection network to derive the latent vectors of I_0 and I_1. The morphed image's latent vector is generated based on a weighted average value, and the resulting vector is employed to generate the final morphed image.

B. Morph Attack Detection

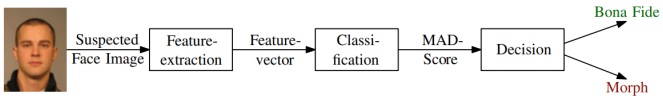
There have been numerous research teams proposing various techniques for detecting morphed images, which can be classified into two categories based on the number of input images: Single Image Morph Attack Detection (SMAD) and Differential Morph Attack Detection (DMAD). The following is an overview of these two methods.

1) Single Image Morph Attack Detection (SMAD): Single image detection, also known as no-reference detection, involves analyzing a single image as input and determining whether it is a morphed image. The detection scheme for a single image is shown in Fig. 2a. Numerous works have focused on the single-image scenario, which is also considered to be more challenging compared to DMAD.

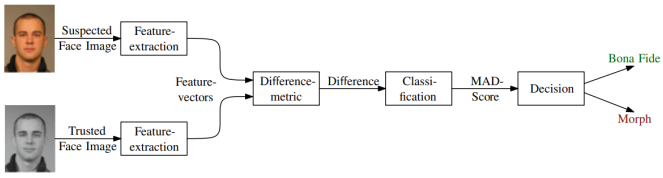
Several studies have employed texture descriptors as a means of detection. This choice is predicated on the conjecture that morphed images may exhibit ghosting artifacts in the hair or neck regions due to imprecise overlapping. Furthermore, automated algorithms commonly generate morphs with half-shade effects on the pupil. In [2], Local Binary Patterns (LBP) [8], and Binarized Statistical Image Features (BSIF) [9] have been used to extract discriminative texture features resulting from morph image processing, followed by the classification task performed by SVM with RBF. Similarly, [10] expanded upon this research by employing high-dimensional LBP feature vectors, yielding promising outcomes. The author in [11] conducted an experiment to replicate real-world scenarios by scanning printed images with two different scanners. The experimental findings demonstrate a higher level of difficulty in detecting printed and scanned morphed image compared to digital morphed images when subjected to the same texture descriptor.

Forensic image analysis approaches focus on detecting traces left by the morphing process on the image, such as sensor pattern noise or inconsistent illumination. [12] implements image source identification using Fourier spectrum of sensor pattern noise, which provides the quantized statistics features for linear SVM training. In a related study [13], the author undertook an analysis of facial highlights and derived an estimation of the light source's position. This information was subsequently utilized to generate a synthesized highlight region, allowing for a comprehensive comparison with the original image.

Deep learning-based methods rely on convolutional neural networks (CNNs) and can be categorized into two types based on their specific implementation. The first type involves using an existing face recognition network to extract features without fine-tuning it [10]. Such research is considered to reduce the risk of overfitting because it does not contain specific morph information in the training task. Secondly, fine-tune the CNN by re-training it on morphed images. For instance, in [14], AlexNet [15] and VGG19 [16] were re-trained to obtain complementary feature to train a Collaborative Representation Classifier. Despite undergoing the process of retraining, the experimental results revealed that the network exhibited a tendency to make judgments solely based on specific artifacts. To address this issue, a method for occluding regions of the



(a) SMAD Scheme



(b) DMAD Scheme

Fig. 2: morphing attack detection scheme [20]

image was proposed in [17].

2) Differential Morph Attack Detection (DMAD): The DMAD will load two images simultaneously, one serving as a reference (i.e., a Suspect Morph), while the other is used as a probe (i.e., a Trusted Live Image). In certain cases, such as in the ABCs scenario, the biological image stored in the eMRTD is the suspect morph, while the live image acquired during the verification procedure is the probe. The accuracy of DMAD can be enhanced by incorporating additional information. The DMAD scheme is depicted in Fig. 2b.

DMAD can be segregated into two research directions. The first involves a direct comparison of the two input images in order to compare the biometric features of the face. This approach extracts the features of the potential morph and the live image, followed by a comparison of these features [3]. It is commonly to observe that SMAD approaches are adapted for application in DMAD scenarios. However, in [18], researchers introduced an innovative detection method solely for DMAD. This method leverages the differences in landmarks between the two images for detection. Another one is to inverse the process of morphing [19]. If the given image is a morph, it is possible to reconstruct the second subject by eliminating the live image (first subject) from the morphed image. In the subsequent comparison, FRSs are likely to reject it due to discrepancies in facial features.

C. Patch-based Works

The utilization of patch-based approach has found applications in diverse research domains. In a recent study [21], a groundbreaking two-stream methodology is introduced for the detection of presentation attacks, combining both local and holistic features. To extract local facial features, a patch-based CNN is employed. The authors indicate that the use of patch-based CNN enables the augmentation of the limited available samples. Instead of rescaling the full image, local patches remain the the discriminative information at the original resolution. Also, one famous patch work—the LBP [8] widely utilized as a texture extractor, can effectively capture the texture characteristics of the patch area. A notable advantage of using LBP is its robustness to illumination variations since each pixel's LBP value is compared with its neighbors. In the research described in [22], a composite model called ConvNet-RBM is employed for facial verification. Each facial image in the dataset is divided into twelve distinct groups, which differ in terms of facial areas and color channels. Within each group, eight modes of patterns are generated using various transformations, such as flipping. The ConvNet model compares corresponding modes from two face images and generates an outcome indicating whether they belong to the same individual. The author highlights that each group focuses on different facial features, resulting in optimized decision-making when utilizing Restricted Boltzmann Machines (RBM) for final classification. A patch-based method for face verification is also suggested by [23]. Resize, area division, color channel alteration, and horizontal flipping are used to enlarge each face's ten patches into a total of 120 images. Then, 60 CNNs process these photos. All the extracted patches features are forwarded to the Joint Bayesian [24] for the last classification. It is empirically shown that the complementary features obtained from these patches significantly contribute to achieving a notably high accuracy score on the LFW dataset [25]. Furthermore, the researchers evaluate the test accuracy by considering the number of patches used.

In classification tasks, where each patch contributes to producing a result, it becomes essential to employ a method for effectively fusing these patch results. The method of combining the classification outputs can be classified into two categories based on the type of classifiers used, as elaborated in [26]. The first category is the hard-level combination, which can be achieved by employing a common mechanism known as the majority vote scheme. The second category is the soft-level combination, where the classifier outputs posteriori probability. Various fusion methods, such as the sum-rule, product rule, max rule, median rule, min rule, and neural networks, can be employed based on this metric level information. In [27], the author proposes MASWOD that can effectively utilize the information from individual classifiers to make the final decision. It is possible to change from hard-level combination to soft-level combination. [28] presents a method that employs the Confusion Matrix and Similarity to obtain the metric level information from classifiers that do not output posteriori probability.

III. Proposed Scheme

To detect morphed images in differential scenarios, we propose the patch-based MAD pipeline. In this section, we provide details on how to use patches for morph detection. Fig. 3 shows the general framework of the approach, which starts with loading two images simultaneously - one is a trusted live image, and the other is a possibly morphed image. Both images are processed to extract patches based on the same rules, and the features of corresponding patches are concatenated to be the input for the classifier. Finally, the score fusion is conducted to enhance the performance.

A. Pre-Processing

In order to ensure that the two patches are located in the same position of the two pictures during feature comparison,

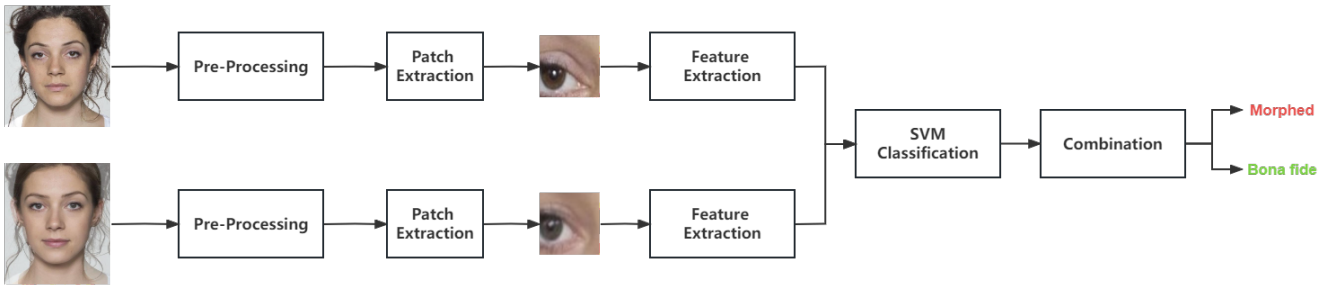


Fig. 3: Patch-based detection pipeline

it is necessary to perform face alignment and face cropping on the input image before patch extraction. Both face alignment and face cropping in our approach are performed using a CNN-based face detector utilizing the Dlib [6]. The method can minimize the impact of various backgrounds when comparing facial patches. It is worth noting that conducting face alignment prior to face cropping effectively prevents the occurrence of uninformative black regions that may arise due to the alignment process. In order to facilitate subsequent segmentation and training procedures, we have resized the cropped faces. Our image resizing strategy was inspired by the dimensions used in [3], specifically 720 × 960 pixels. However, considering that our experiments solely involve facial parts, we have adjusted the cropped images to a size of 480 × 576 pixels accordingly.

B. Patch Extraction

After having undergone identical pre-processing, the two face images that are of equal size are sent to the patch cutter. The patch processor should be configured to divide both images into an equal number of patches using the built-in method. The patches of the two pictures can be matched one by one in a comparative analysis.

The patch extraction method we employed for experimentation is complete image segmentation, which is known for its simplicity and ease of implementation. This technique involves dividing the image directly into uniform patches. The feasibility of this approach has been validated in previous works such as [29] and [30]. In our work, we referred to the patch size utilized in [21] and divided each image into a total of 30 patches, each measuring 96 × 96 pixels.

C. Feature Extraction

The feature extraction technique utilized in SMAD can also be employed in DMAD. The detailed methodology for texture descriptors, forensic image analysis, and deep learning can be found in the second section. In theory, these feature extraction methods can be used in the patch context, but we focus on deep learning in the paper. Based on previous research, the neural network architecture shown in Table I has demonstrated viability in detecting morphs across the entire face [14] [31]. These architectures are therefore strong contenders for use in extracting patch-based features for analysis. In Experiment 1, we employed three network architectures to assess their performance. However, in subsequent experiments, we exclusively utilized the network architecture that demonstrated the highest performance in Experiment 1. Before forwarding the features into the classifier, two feature vectors extracted from corresponding patches must be concatenated.

TABLE I: Network architectures

Dataset	Network
ImageNet [32]	Se_ResNet50 [33]
	ResNet50 [34]
	VGG-19 [16]

D. Classification

With the additional information contained in the trusted live patch, it is feasible to assess differences between those two feature vectors. In the study of [3] [18], Random Forest, AdaBoost, Gradient Boosting, SVM were compared. Support Vector Machines (SVMs) exhibited commendable results among various machine learning-based classifiers.

E. Combination

To determine the final decision, we employed three distinct methods in our experiment: majority voting, score averaging, and machine learning. In Experiment 1, we applied all three methods to assess their effectiveness. However, in subsequent experiments, we focused on the fusion mechanism that exhibited superior performance in Experiment 1.

The majority vote mechanism comprises two distinct stages (see 1 and 2). In the first stage, this classifier assigns a score s_i to each patch. These scores are subsequently compared to a predefined threshold t, enabling the determination of the individual patch's judgment result. Next, the votes V_i of individual patches within each image are summed, resulting in the calculation of the final score, denoted as S. Subsequently, S is compared against a predefined threshold T, in order to arrive at the ultimate decision for the entire image.

$$V_i = \begin{cases} 0, & s_i < t \\ 1, & s_i \geq t \end{cases} \tag{1}$$

$$D = \begin{cases} genuine, & S = \sum_{i=1}^{n} V_i < T \\ impostor, & S = \sum_{i=1}^{n} V_i \geq T \end{cases} \quad (2)$$

The second score fusion mechanism entails employing the average score of patches within an image to determine its classification. If the computed average score surpasses a predefined threshold value T, the image in the reference is categorized as "morph," as demonstrated in 3.

$$D = \begin{cases} genuine, & S = \frac{1}{n}\sum_{i=1}^{n} s_i < T \\ impostor, & S = \frac{1}{n}\sum_{i=1}^{n} s_i \geq T \end{cases} \quad (3)$$

The machine learning approach involves constructing a novel feature vector by utilizing the scores of the 30 patches within the image. This newly formed feature vector is then inputted into an SVM classifier to generate the final score S. Similarly, if the obtained score exceeds the threshold value T, the final decision D is designated as "impostor".

IV. EXPERIMENTS AND RESULTS

This section describes the employed datasets, the involved morph algorithms, and the evaluation metrics. Subsequently, we conducted a comparison of various combinations of feature extraction network architectures and patch score fusion methods. The most effective combination was employed in extensive experiments to explore the influence of training data on the performance of the system.

A. Creation of MAD Dataset

To investigate DMAD, it is necessary to consider not only the reference image (bona fide and morphed image), but also an additional probe, such as live images captured at the eGate. The reference images adhere to stringent criteria concerning the environment, lighting conditions, and other factors. For instance, in accordance with ICAO regulations, the distance between the eyes in reference images should be no less than 90 pixels. However, the probe images are obtained within a semi-controlled setting, less constraints on lighting, posture, facial expression, and other variables.

The experiment employed three distinct datasets, namely FRGC [35], PUT [36], and FRLL [37].

1) FRGC: FRGC is a widely utilized dataset for MAD tasks, offering high-resolution reference and probe images captured in diverse environments. For each identity in our experiment, two passport-quality images were selected, one for bona fide purposes and the other for generating morph images. Each reference image had a corresponding probe image. The dataset was partitioned into non-overlapping identities, with 121 subjects allocated to the training set, 41 subjects to the validation set, and 41 subjects to the test set.

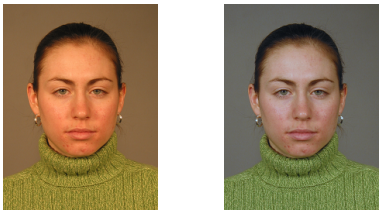
2) PUT: The PUT dataset, with a limited number of identities, comprised images captured under controlled conditions, exhibiting variations in facial appearance due to different head orientations. If there were two images of an identity with the head in a forward position, one was used as a bona fide image, and the other for morph generation. Regarding probes, efforts were made to select images with slightly tilted heads. Ultimately, only 88 identities fulfilled the criteria, and all of them were used for training.

3) FRLL: In contrast to the previous two datasets, we directly employed the publicly available FRLL-Morph dataset [38], which is based on FRLL. The morphed images in the FRLL-Morph dataset are significantly compressed, resulting in a much lower resolution compared to morphed images from other sets. All subjects in this dataset were used for cross-dataset performance evaluation.

Six different algorithms were employed for morph generation in experiments.

- OpenCV: This open-source morphing tool utilizes Dlib for landmark detection [3]. It generates Delaunay triangles for wrapping and blending. Notably, Opencv adds additional keypoints in areas such as shoulders and image edges for morph generation. However, due to the absence of keypoints outside the facial region, ghosting artifacts can be observed.
- Webmorph: This online landmark-based generation tool [39] is specifically designed for FRLL and requires 189 landmark points, a criterion fulfilled only by FRLL. Nevertheless, noticeable ghosting artifacts can still be observed, particularly in areas like hair.
- Combined Morphs: Proposed in [40], this novel landmark-based morphing method employs Dlib landmark detection. The algorithm effectively avoids ghosting artifacts by splicing the synthesized region of the face into a wrapped face. Furthermore, Poisson image editing is employed to enhance the natural appearance of the morph images.
- UBO: This method, utilized in the paper by the University of Bologna [41], is similar to Combined Morphs as it involves the splicing of the synthesized facial region with the external region, but it employs weighted blending at the edges.
- UTW: It is a morphing algorithm developed at the University of Twente [42]. Similar to Combined Morphs, it incorporates background replacement but uses the STASM [43] landmark detector instead.
- STYLEGAN2 [44]: This GAN-based algorithm generates highly natural images with minimal visible artifacts. It employs a pre-trained projection model to obtain latent vectors for the input images, which are then interpolated to generate the morphed image.
- MIPGAN [45]: It is an improved algorithm based on STYLEGAN. The newly introduced loss function ensures the identity preservation of the morphed image.

[3] https://learnopencv.com/face-morph-using-opencv-cpp-python/



(a) Original (b) Normalized

Fig. 4: White Balance

TABLE II: Combination Elements

Dataset	Action	Method	Number
FRGC	original	UTW	1
		STYLEGAN2	2
	normalization	UTW	3
		STYLEGAN2	4
PUT	original	STYLEGAN2	5

We selected two distinct methods, a landmark-based approach UTW and a GAN-based method STYLEGAN2, in the training phase. Additionally, we utilized a tool called whitebalance [4] to modify the overall style of the FRGC dataset, as depicted in Fig. 4. Each row in Table II illustrates various elements for constructing different training sets. For ease of reference, we assigned the numerical identifier to each one. Table III presents the diverse settings utilized for each experiment. All experiments will perform tests utilizing the FRGC dataset (referred to as FRGC-Morph) and FRLL-Morph dataset, with detailed particulars provided in Table IV.

B. Performance Evaluation

In all experiments, the assessment of outcomes will rely on the evaluation metrics of APCER (Attack Presentation Classification Error Rate) and BPCER (Bona Fide Presentation Classification Error Rate), which are commonly employed in the context of morphing attack detection. These metrics are

[4] http://www.fmwconcepts.com/imagemagick/whitebalance/index.php

TABLE III: Experiment settings

Experiment	Train
1	1+2
2	1+2+3+4
3	1+5
4	1
5	5
6	2
7	4

TABLE IV: Test set

Dataset	Morphed Images		Bona fide
FRGC-Morph	UTW	27	80
	STYLEGAN2	13	
	UBO	27	
	MIPGAN	14	
FRLL-Morph	Combined Morph	74	85
	Webmorph	81	
	OpenCV	81	
	STYLEGAN2	82	

TABLE V: Images in Training Set - Experiment1

Source Dataset	Morphed Images		Bona fide
FRGC	UTW	131	192
	STYLEGAN2	73	

defined in ISO/IEC 30107-3 [46]. APCER quantifies the ratio of morph attack presentations that are inaccurately classified as bona fide presentations, while BPCER measures the proportion of bona fide presentations that are inaccurately classified as morph attack presentations. The DET (Detection Error Trade-off) curve is constructed by plotting the values of APCER and BPCER with varying the threshold. Moreover, the detection model's accuracy is assessed using D-EER (Detection Error Equal Rate), which quantifies the error rate when the APCER and BPCER are the same.

C. Experiment 1 - Training on the Original FRGC with UTW and STYLEGAN2

In Experiment 1, we assessed the model's ability to generalize when trained exclusively on the FRGC dataset, while utilizing both the UTW and STYLEGAN2 morphing generation algorithms. Furthermore, we combined the three feature extraction networks with three patch fusion methods mentioned in Section 3. The most effective combination was chosen as the feature extractor and score combination technique for subsequent experiments. The training data information can be found in Table V.

For the sake of analytical convenience, we conducted separate tests on the images generated by different morph algorithms in the FRLL-Morph dataset. The detection performance is presented in the Table VI.

When employing Se_Resnet50 as the feature extractor, it is evident that regardless of the morph generation algorithm, the detector can effectively distinguish between bona fide and morphed images as long as the test images are also from FRGC. Conversely, in the context of cross-dataset , it is observed that morphed images produced by STYLEGAN2 are the most difficult to detect. Although cross-dataset testing has traditionally posed difficulties for morph detection due to varying dataset characteristics, our methods still exhibit generalized capability for morphed images generated by Combined Morph, OpenCV, and Webmorph. Notably, despite the noticeable presence of artifacts outside the facial region in OpenCV-generated images, the patch-based detector primarily relies on the cropped facial region, which remains challenging to discern. However, the scenario changes significantly when Resnet50 is utilized for extracting features. Regardless of the employed scoring fusion method, the detector can only accurately distinguish images from the FRGC dataset. With the exception of morphed images generated by OpenCV, the model exhibits an inability to effectively process the remaining images from the FRLL-Morph dataset. This discrepancy indicates that the weights assigned to the channels have a substantial impact on the model's performance. When the VGG19 network architecture is employed as the feature extractor, even samples from the FRGC dataset can result in classification errors. Morphed images generated by STYLEGAN2 remain challenging to

TABLE VI: Detection performance for different Morph algorithms

Experiment	**Feature Extractor**	**Fusion Methods**	**D-EER (in %)**				
			Combined Morphs	OpenCV	STYLEGAN2	Webmorph	FRGC-Morph
1	Se_Resnet50	Without Fusion	15.3	26.2	54.5	20.0	5.8
		Average	1.2	12.1	55.7	3.6	0.0
		Vote	1.9	12.7	52.6	4.2	0.0
		Machine Learning	1.2	12.1	55.7	3.6	0.0
	Resnet50	Without Fusion	47.8	36.6	47.3	51.2	9.8
		Average	46.6	19.3	45.5	56.7	0.0
		Vote	52.6	21.1	45.5	62.0	0.0
		Machine Learning	45.9	21.1	40.7	54.2	0.0
	VGG19	Without Fusion	8.4	26.6	60.5	9.0	9.3
		Average	0.7	18.1	62.9	4.2	0.6
		Vote	1.2	19.9	65.2	2.4	2.5
		Machine Learning	1.9	15.7	62.9	3.6	1.2
2	Se_Resnet50	Average	46.7	12.7	41.3	53.0	0.0
3	Se_Resnet50	Average	24.5	17.5	29.9	30.7	0.0
4	Se_Resnet50	Average	70.4	48.8	52.7	66.9	10.6
5	Se_Resnet50	Average	27.8	29.5	22.7	39.1	27.9
6	Se_Resnet50	Average	0.0	4.8	37.7	0.0	32.9
7	Se_Resnet50	Average	6.3	3.0	37.7	9.6	24.2

detect. However, regardless of the feature extractor utilized, when the trained model demonstrates effectiveness on the test set, the accuracy of detection can be enhanced through the implementation of a fusion strategy.

The decline in performance during cross-dataset testing could be attributed to several factors. In contrast to FRGC-Morph, the morphed images contained within the FRLL-Morph dataset have undergone a compression process that eliminates many morphing traces, thus leading to a decline in the performance of the detection system. Additionally, observable visual discrepancies between the FRGC and FRLL datasets are evident. In light of this observation, we will proceed with subsequent experiments to evaluate the performance variations of models trained on distinct training sets. To determine the optimal combinations, we evaluated three feature extraction networks and three score fusion mechanisms by calculating the average value of all D-EERs except for STYLEGAN2 (FRLL-Morph) column. Our findings indicate that the combinations of Se_Resnet50 with average and Se_Resnet50 with machine learning are the optimal choices. We selected Se_Resnet50 with average fusion for the subsequent experiments.

D. Experiment 2 - Training on the Original and Normalized FRGC with UTW and STYLEGAN2

In Experiment 2, in order to achieve a visual similarity with the FRLL dataset, a technique known as "white balance" was employed to modify the visual appearance of images from the original dataset (FRGC). The details of the training set in Experiment 2 are presented in Table VII. The model's detection performance is depicted by the DET curve shown in Figure 5.

The experimental results indicate improvements in the model's performance when detecting images generated by STYLEGAN2, yielding D-EERs of 41.3%. The finding suggests that altering background style might have the potential to enhance the effectiveness of the detector when dealing with images generated by this morph tools. However, for morphed images generated by other algorithms, it appears

TABLE VII: Images in Training Set - Experiment2

Source Dataset	Morphed Images		Bona fide
FRGC	UTW	73	146
	STYLEGAN2	73	
FRGC (Normalization)	UTW	73	146
	STYLEGAN2	73	

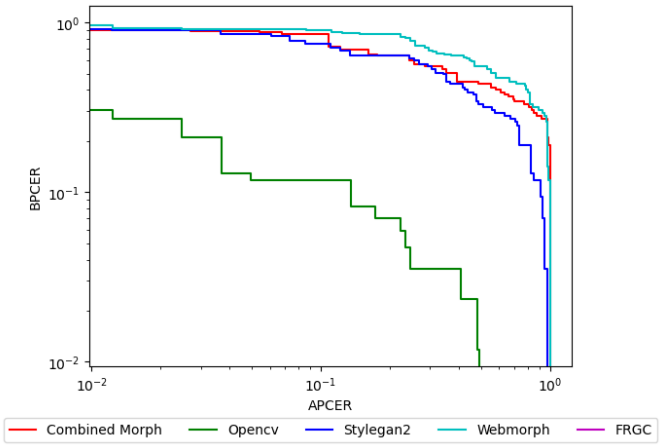


Fig. 5: DET curve of Experiment 2

that the application of white balance has a detrimental effect, leading to a decrease in the patch-based detection system's performance. Nonetheless, the system can still accurately differentiate morphed images from the FRGC dataset.

E. Experiment 3 - Training on the PUT with UTW and STYLEGAN2

In Experiment 3, a novel dataset called PUT was introduced into the training set, which differs significantly from the FRGC dataset in terms of background lighting. The images in PUT exhibit a predominantly whitish color, resembling the visual style of FRLL. The details of the images utilized in Experiment 3 are presented in Table VIII. Figure 6 illustrates notable changes in the detection results compared to the experiment 1. These changes primarily manifest in the model

TABLE VIII: Images in Training Set - Experiment3

Source Dataset	Morphed Images		Bona fide
FRGC	UTW	74	74
PUT	STYLEGAN2	76	76

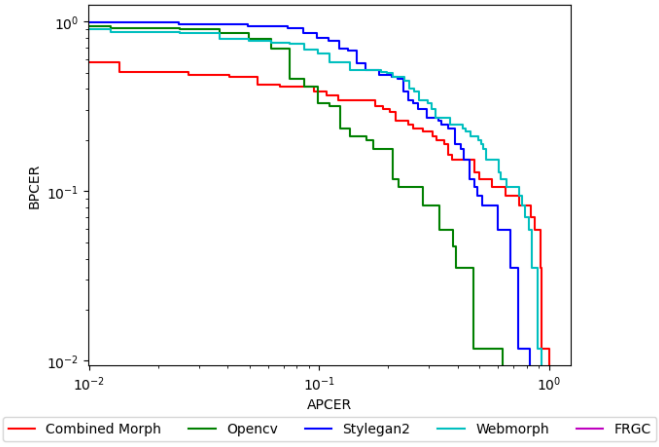


Fig. 6: DET curve of Experiment 3

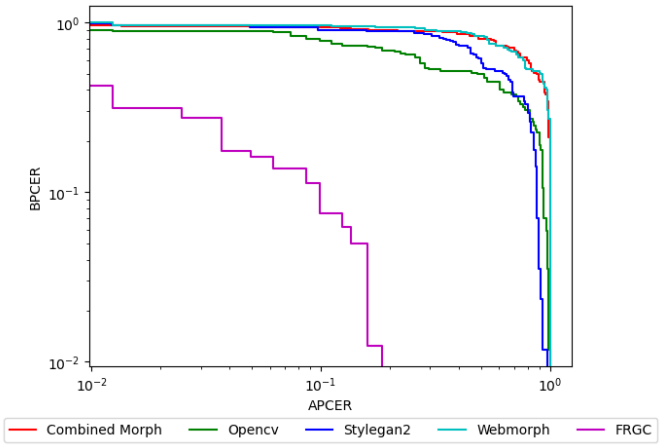


Fig. 7: DET curve of Experiment 4

TABLE X: Images in Training Set - Experiment5

Source Dataset	Morphed Images		Bona fide
PUT	STYLEGAN2	76	76

exhibiting higher error rates when presented with images generated by Combined Morphs, OpenCV, and Webmorph. This signifies that images produced by STYLEGAN2 using the FRGC dataset are important for the detection of landmark-based morph images in the FRLL-Morph test set. However, encouragingly, the performance of the new model has significantly improved for images generated by STYLEGAN2 in FRLL-Morph.

F. Experiment 4 - Training on the FRGC with UTW

Experiment 4 reveals the distinct performance of a model trained on UTW-generated images using FRGC when applied to diverse test sets. The specific details of the training set can be found in Table IX. Notably, as depicted in Figure 7, during cross-dataset testing, the model's generalization capability diminishes, resulting in significantly high D-EER when faced with all morphed images. Although the D-EER on FRGC-Morph reaches 10.6%, the primary errors originate from images generated by STYLEGAN2. The model demonstrates proficiency in detecting images generated by UTW and MIPGAN algorithms, but it encounters classification errors when confronted with images generated by UBO. These findings indicate that landmark-based and GAN-based morph images may possess limited shared features.

G. Experiment 5 - Training on the PUT with STYLEGAN2

Experiment 5 was dedicated exclusively to training the model using samples generated by STYLEGAN2 based on PUT dataset. The quantities of morphed and bona fide images are presented in Table X. The experimental results depicted in Figure 8 demonstrate that the model exhibits a certain degree of detection effectiveness for STYLEGAN2-generated images in FRLL-Morph. The D-EER for STYLEGAN2 (22.7%) is even lower than the D-EER observed in Experiment 3 (29.9%). This implies that a model trained exclusively on images created by STYLEGAN2 can exhibit improved adaptability when confronted with STYLEGAN2-generated images from different datasets. Additionally, it shows the ability to perceive landmark-based morphed images, suggesting that both GAN-based morphs and landmark-based morphs share similar characteristics to some extent. Furthermore, it is apparent that the detection system exhibits significant errors when confronted with images from the FRGC dataset. However, further analysis reveals that this is primarily attributable to the presence of landmark-based images in the FRGC test set. When encountering images generated by STYLEGAN2 and MIPGAN algorithms from the FRGC-Morph, the model can still accurately distinguish them.

TABLE IX: Images in Training Set - Experiment4

Source Dataset	Morphed Images		Bona fide
FRGC	UTW	131	131

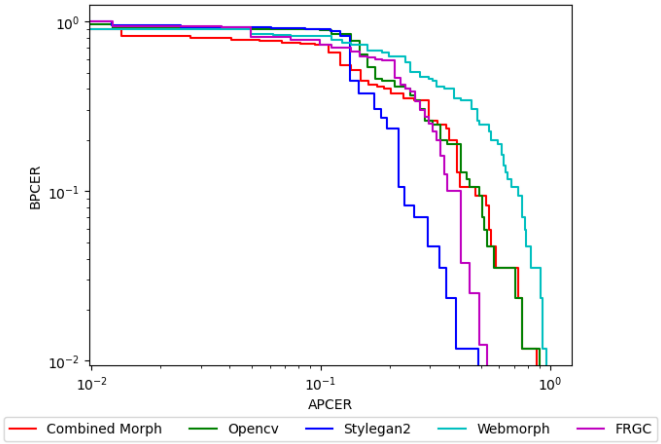


Fig. 8: DET curve of Experiment 5

TABLE XI: Images in Training Set - Experiment6

Source Dataset	Morphed Images		Bona fide
FRGC	STYLEGAN2	74	75

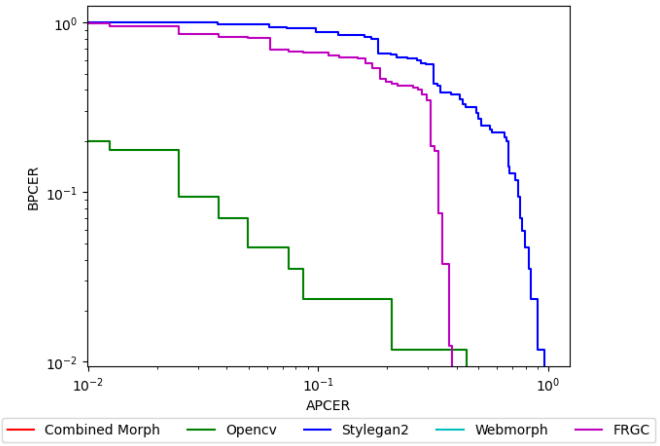


Fig. 9: DET curve of Experiment 6

TABLE XII: Images in Training Set - Experiment7

Source Dataset	Morphed Images		Bona fide
FRGC (Normalization)	STYLEGAN2	74	75

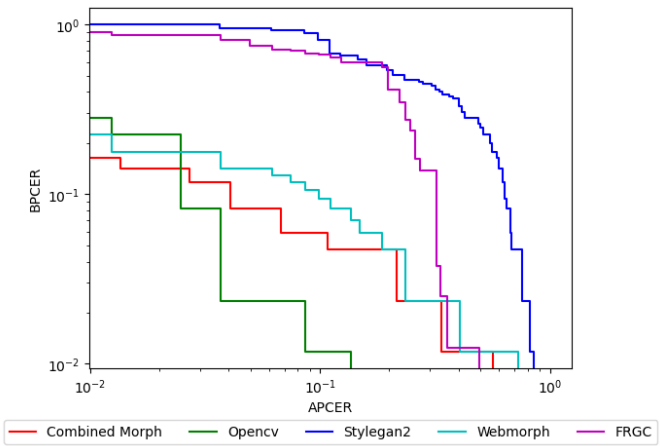


Fig. 10: DET curve of Experiment 7

H. Experiment 6 - Training on the Original FRGC with STYLEGAN2

Experiment 6 aims to investigate the detection performance of the generated model by exclusively utilizing morph images generated by STYLEGAN2 based on the FRGC. The experiment's training set information and experimental results are presented in Table XI and Figure 9, respectively. Upon testing on FRGC-Morph, the model demonstrates a D-EER of 32.9%, solely attributed to landmark-based morph images within the test set. In contrast to the D-EER value of 55.1% observed in Experiment 1, the new model exhibits a lower D-EER of 37.7% when confronted with images generated by STYLEGAN2 in the FRLL-Morph. This observation provides additional support for the inference made in Experiment 5. A model solely trained on images produced by STYLEGAN2 showed the better generalized capability when presented with STYLEGAN2-generated images originating from diverse datasets. Nevertheless, the figure exhibits an approximate 15% elevation when contrasted with the D-EER from Experiment 5. This discrepancy is likely attributed to the significant disparities in background style between the FRGC dataset and FRLL-Morph dataset.

I. Experiment 7 - Training on the Normalized FRGC with STYLEGAN2

With the aim of delving more directly into the potential positive impact of white balance on the detection of STYLEGAN2-generated Morph images during cross-dataset testing, the present experiment employed normalized images for training the detection model. Comprehensive training specifics are shown in Table XII. The performance of detection across various morph algorithms is in Figure 10. In comparison to Experiment 6, the normalization strategy resulted in a slight reduction in detection error rates for OpenCV-generated images and FRGC-Morph. Similarly, when examined using the FRGC-Morph dataset, the model exhibited a D-EER of 24.2%, exclusively ascribed to landmark-based morph images within the test set. However, the D-EER for images generated by STYLEGAN2 remained consistent with the values observed in Experiment 6. This unexpected difference from the initial inference highlights the need for further investigation to understand potential factors.

V. Conclusion

In summary, our experimental findings provide evidence that utilizing patches for morphing attack detection is a viable approach, and the application of the scores fusion technique notably enhances the detection accuracy. By employing the Se_resnet50 feature extractor in conjunction with the average fusion method, the proposed scheme showcases flawless performance on the FRGC-Morph test set. Additionally, despite the compression applied to the morph images in FRLL-Morph, it also demonstrates satisfactory detection results on the FRLL-Morph (with the exception of STYLEGAN2). Through conducting extensive experiments, we note that the training data significantly influences the performance of the final detection system. Despite the distinct generation methods employed for GAN-based morphs and landmark-based morphs, we discovered that when the training set exclusively comprises morphed images generated by STYLEGAN2, the resulting model can still exhibit some level of recognition for landmark-based morphs. Futhermore, a model exclusively trained on images generated by STYLEGAN2 demonstrated superior generalizability when confronted with STYLEGAN2-generated images originating from different datasets.

Nevertheless, the current findings only indicate that enhancing the resemblance between the backgrounds of training and test images holds the potential to better detect morphs generated by STYLEGAN2. Nonetheless, this pattern may not always manifest clearly. Further experimentation is required to analyze it. Moreover, since a substantial majority of morphed images within the FRLL-Morph dataset have

undergone significant compression, introducing images with varying resolutions during the training process could overcome it. In order to achieve a higher level of detection performance, additional exploration for the quantity and size of patches , as well as the approach used for segmenting is required.

REFERENCES

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

[2] U. Scherhag, C. Rathgeb, and C. Busch, "Towards detection of morphed face images in electronic travel documents," in *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*. IEEE, 2018, pp. 187–192.

[3] U. Scherhag, C. Rathgeb, J. Merkle, and C. Busch, "Deep face representations for differential morphing attack detection," *IEEE transactions on information forensics and security*, vol. 15, pp. 3625–3639, 2020.

[4] M. Ferrara and A. Franco, "Morph creation and vulnerability of face recognition systems to morphing," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing Cham, 2022, pp. 117–137.

[5] D. J. Robertson, R. S. Kramer, and A. M. Burton, "Fraudulent id using face morphs: Experiments on human and automatic recognition," *PLoS One*, vol. 12, no. 3, p. e0173319, 2017.

[6] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[7] B. Delaunay, "Sur la sphère vide," *Bulletin de l'Academie des Sciences de l'URSS. Classe des sciences mathematiques et na*, vol. 1934, no. 6, pp. 793–800, 1934.

[8] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[9] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE, 2012, pp. 1363–1366.

[10] L. Wandzik, G. Kaeding, and R. V. Garcia, "Morphing detection using a general-purpose face recognition system," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1012–1016.

[11] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the vulnerability of face recognition systems towards morphed face attacks," in *2017 5th international workshop on biometrics and forensics (IWBF)*. IEEE, 2017, pp. 1–6.

[12] L.-B. Zhang, F. Peng, and M. Long, "Face morphing detection using fourier spectrum of sensor pattern noise," in *2018 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2018, pp. 1–6.

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

[14] K. Raja, S. Venkatesh, R. Christoph Busch *et al.*, "Transferable deep-cnn features for detecting digital and print-scanned morphed face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 10–18.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Seibold, W. Samek, A. Hilsmann, and P. Eisert, "Accurate and robust neural networks for security related applications exampled by face morphing attacks," *arXiv preprint arXiv:1806.04265*, 2018.

[18] U. Scherhag, D. Budhrani, M. Gomez-Barrero, and C. Busch, "Detecting morphed face images using facial landmarks," in *Image and Signal Processing: 8th International Conference, ICISP 2018, Cherbourg, France, July 2-4, 2018, Proceedings 8*. Springer, 2018, pp. 444–452.

[19] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, "Border control morphing attack detection with a convolutional neural network de-morphing approach," *IEEE Access*, vol. 8, pp. 92 301–92 313, 2020.

[20] U. Scherhag, C. Rathgeb, and C. Busch, "Face morphing attack detection methods," in *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*. Springer International Publishing Cham, 2022, pp. 331–349.

[21] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 319–328.

[22] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1489–1496.

[23] ——, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1891–1898.

[24] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part III 12*. Springer, 2012, pp. 566–579.

[25] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[26] M. Mohandes, M. Deriche, and S. O. Aliyu, "Classifiers combination techniques: A comprehensive review," *IEEE Access*, vol. 6, pp. 19 626–19 639, 2018.

[27] J. Zhang, L. Cheng, and J. Ma, "A new multiple classifiers combination algorithm," in *First International Multi-Symposiums on Computer and Computational Sciences (IMSCCS'06)*, vol. 2. IEEE, 2006, pp. 287–291.

[28] Y.-D. Lan and L. Gao, "A new model of combining multiple classifiers based on neural network," in *2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies*. IEEE, 2013, pp. 154–159.

[29] H. R. Kanan, K. Faez, and Y. Gao, "Face recognition using adaptively weighted patch pzm array from a single exemplar image per person," *Pattern Recognition*, vol. 41, no. 12, pp. 3799–3812, 2008.

[30] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[31] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni, "A double siamese framework for differential morphing attack detection," *Sensors*, vol. 21, no. 10, p. 3466, 2021.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

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

[36] A. Kasinski, A. Florek, and A. Schmidt, "The put face database," *Image Processing and Communications*, vol. 13, no. 3-4, pp. 59–64, 2008.

[37] L. DeBruine and B. Jones, "Face research lab london set," May 2017.

[38] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," *arXiv preprint arXiv:2012.05344*, 2020.

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

[40] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, no. 4, pp. 325–332, 2018.

[41] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 4, pp. 1008–1017, 2017.

[42] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos, F. de Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. K. Venkatesh *et al.*, "Morphing attack detection-database, evaluation platform, and benchmarking," *IEEE transactions on information forensics and security*, vol. 16, pp. 4336–4351, 2020.

[43] S. Milborrow and F. Nicolls, "Active shape models with sift descriptors and mars," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2. IEEE, 2014, pp. 380–387.

[44] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[45] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, "Mipgan—generating strong and high quality morphing attacks using identity prior driven gan," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 365–383, 2021.
[46] I. Standard, "Information technology–biometric presentation attack detection–part 3: testing and reporting," *International Organization for Standardization: Geneva, Switzerland*, 2017.