2022

Generation of High Performing Morph Datasets

Kelsey Lynn O'Haire
klo0003@mix.wvu.edu

Generation of High Performing Morph Datasets

Kelsey L. O'Haire

Thesis submitted to the Benjamin M. Statler College of Engineering and Mineral Resources at West Virginia University in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

Nasser M. Nasrabadi, PhD., Chair,

Jeremy Dawson, PhD.,

Omid Dehzangi, PhD.

Morgantown, West Virginia

2022

Keywords: Face Recognition, Morph Attacks, Morph Generation, Wavelet Sub-bands

Abstract

Generation of High Performing Morph Datasets

Kelsey L. O'Haire

Facial recognition systems play a vital role in our everyday lives. We rely on this technology from menial tasks to issues as vital as national security. While strides have been made over the past ten years to improve facial recognition systems, morphed face images are a viable threat to the reliability of these systems. Morphed images are generated by combining the face images of two subjects. The resulting morphed face shares the likeness of the contributing subjects, confusing both humans and face verification algorithms. This vulnerability has grave consequences for facial recognition systems used on international borders or for law enforcement purposes. To detect these morph images, high-quality data must be generated to improve deep morph detectors.

In this work, high-quality morph images are generated to fool these deep morph detection algorithms. This work creates some of the most challenging large-scale morphed datasets to date. This is done in three ways. First, rather than utilizing typical datasets used for face morphing found in literature, we generate morphed data from underrepresented groups of individuals to further increase the difficulty of morphs. Second, we generate morph subjects using a wavelet decomposition blending technique to generate morph images that may perform better than typical landmark morphs while creating morph images that may appear different to detectors than what is seen in literature. Third, we apply adversarial perturbation to the morph images to further increase their attack capability on morph detectors. Using these techniques, the generated morph datasets are highly successful at fooling facial recognition systems into erroneously classifying a morph as a bona fide subject.

Acknowledgments

This work was completed with the help of many people. First, I would like to thank my advisor, Dr. Nasrabadi, for his wisdom, insight and confidence in me over the last two years. Next, I would like to thank my committee members, Dr. Dawson and Dr. Dehzangi, for their time and support. I thank Sobhan Soleymani, for without his guidance, this work would not have been possible. Additionally, I need to thank all of my peers in Dr. Nasrabadi's research lab who have become both my friends and mentors.

Last, I want thank all of my friends and family for their unwavering support and for being the rock in my life. Specifically, my Mom and Dad who have supported me endlessly. My sister, Emily, who I look forward to seeing her own success at WVU. And finally, my boyfriend, Jeremy Keith, who is my best friend and biggest cheerleader.

Publications

1. Kelsey O'Haire, Sobhan Soleymani, Baaria Chaudhary, Poorya Aghdaie, Jeremy Dawson, and Nasser M. Nasrabadi. "Adversarially Perturbed Wavelet-based Morphed Face Generation." In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pp. 01-05. IEEE, 2021.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biometric Systems

Biometrics are physical attributes that are used to automatically link a person to their identity [1]. The four stages of a biometric system are enrollment, template creation, identification, and verification [2]. Biometrics act as a key to unlocking the identity of an individual. Further, biometrics are common at border control checkpoints and are considered a national security necessity [3]. Specifically, the face as a biometric is commonly used in government identification documents [1, 2, 3]. The face is a convenient biometric because it is non-intrusive and easy to verify in person [3, 4, 5].

Due to the advent of Deep Learning (DL) technology, facial recognition systems (FRS) have been improved significantly over the past ten years. The first DL based facial recognition model was Facebook's Deepface [6], which was trained on over four million images from four thousand subjects [7]. Deepface has an accuracy of 97.35% on the Labeled Faces in the Wild dataset and has an error rate more than 25% lower than the next-best performing algorithm at the time [7]. NIST observed that facial recognition technology became twenty times better at accurately matching individuals between the years of 2014 to 2018,

with only 0.2% of matches failed in 2018 [8].

As technology improves, it permeates into daily life. We see facial recognition being used in public spaces like airports, grocery stores, and schools, but we also bring it home with us in our our cell phones and laptops [7]. Today, 15% to 20% of financial institutions in the U.S. utilize facial recognition to verify clients. With so much of our daily lives reliant on the reliability of these systems, it is vital that facial recognition systems be thoroughly tested for vulnerabilities and short-comings. These vulnerabilities typically manifest from the improper training of DL models. The accuracy of DL models are constrained to the data that they are trained on. Without proper training data, facial recognition systems will fail when they encounter subject data that is not typical with what the system is trained on. For example, Buolamwini *et al.* showed that commercial facial recognition systems have error rates up to 30% higher on subjects with dark skin compared to subjects with light skin in a gender classification scenario. Clearly, underrepresented scenarios in training data lead to a drop in facial recognition system performance.

1.2 Motivation

While FRS are a security necessity, they are vulnerable to attacks in the enrollment stage. If an enrolled passport photo resembles multiple people, the passport can be shared between the look-alikes. Morphed images are created by combining face images from two or more individuals, creating a new ambiguous face which possesses similarities between the bona fide identities. As morphing technology becomes more accessible, anyone can create high-quality morphed images with little to no technical background, highlighting the need for more challenging morphed image datasets for training face morph detectors.

Issues arise when bad actors intentionally try to spoof these systems. False positives in biometric systems allow for bad actors to pass through a the system under the alias of

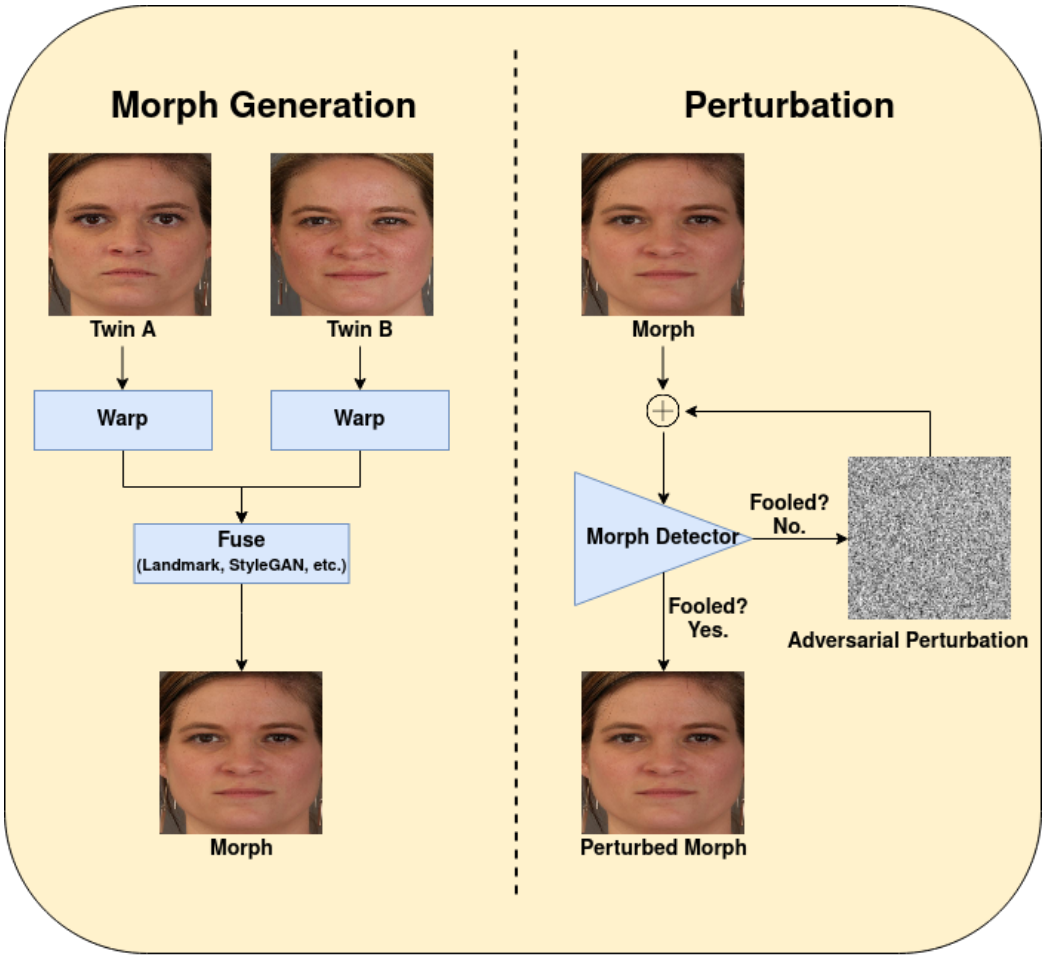

Figure 1.1: (left) Morph generation overview. (right) Perturbation is added to morphed images to further increase morph detection difficulty.

another person's identity. False positives occur when two people look alike and the FRS authenticates the look-alikes as one another. Morph images are faces made with the express purpose of fooling a FRS into falsely verifying one individual as another.

The goal of this work is to create the most challenging morph datasets to date. We do this in two ways, first, we leverage edge-cases scenarios of FRS in order to improve the attack capabilities of our morphs. For example, we generate morphs using both a juvenile and a twin dataset. Additionally, we adversarially perturb the images in order to make the attack quality of the images more challenging for FRS. Second, we utilize a new way

of morphing images in the wavelet domain in order to further improve attack capabilities and potentially fool morph detectors that are trained to detect specific morphing types. An example of the morphing pipeline can be seen in Figure 1.1.

Our morphed images are generated with three separate methodologies, landmark [9, 10], Generative Adversarial Network (GAN)-based models [11, 12], and wavelet-based morphing [13]. For landmark morphing, critical points of the two input subjects are averaged together to create common landmarks. The images are then warped towards these common landmarks and blended to create the morphed image. The GAN-based approach uses latent vectors of input images which are then linearly combined, resulting in minimal artifacts and producing high-quality morphs [12, 14]. Further, we introduce a new morph generation method utilizing the Discrete Wavelet Transform (DWT), where the input images are warped and then subsequently blended using their wavelet decomposed sub-bands. Finally, we apply adversarial perturbation to these morphed image in order to further increase verification difficulty by fooling a classifier into mislabeling the examples into an erroneous class.

1.3 Contribution of Thesis

In this thesis, we push the bounds of morph generation into creating the most difficult morph images possible. Morph dataset generation is vital for the improvement of morph detectors, especially for Deep Learning models that require large amounts of data. Currently, there is a lack of high-quality morph datasets found in literature. Without this data, there is an obvious national security concern, with the potential of bad actors to either sneak under the alias of a morph image, or for human trafficking which would allow for vulnerable people to be snuck out of a country. Thus, the contribution in this thesis is as follows:

- Create a new high-quality morph dataset using the FERET [15] and FRGC [16] datasets which is free of shadowing using both the landmark and wavelet-based morphing techniques.
- Generate a morph dataset made entirely of juvenile subjects whose ages range from toddler to late teenager.
- Create a morph dataset using Identical Twins to create the worst case scenario for morphed images.
- Utilize morph generation method utilizing the Discrete Wavelet Transform (DWT), where the input images are warped and then subsequently blended using their wavelet decomposed sub-bands. Our work is the first that leverages the spatial-frequency wavelet domain to create high quality morphs.
- After morphing, a visually indistinguishable amount of adversarial perturbation is applied to further increase the difficulty of detecting the morphed face images. The generated high-quality morphed images have no obvious signs of tampering and show high similarity to individuals combined in the morphing process.

1.4 Organization of Thesis

This work is comprised of three main sections after this chapter. Chapter 2 discusses the literature generation process of Landmark, Wavelet, and StyleGAN2 morphs. The benefits and short-comings of these methods are discussed. Additionally, this chapter contains information regarding the adversarial perturbation procedure utilized, as well as a section dedicated to morph detection. Chapter 3 discusses the generated datasets with their respective analysis. We discuss three generated datasets, the Baseline Wavelet dataset, Child

dataset, and the Twin dataset. The Baseline Wavelet dataset is evaluated first, then the Twin morph dataset, and lastly the Child morph dataset. Data is analyzed in terms of error rates, classification rates, and MMPMR. Last, in Chapter 4, we will discuss the conclusion and next steps for the generated morph datasets.

Chapter 2

Morph Generation

Ferrara *et al.* [17] are the first to investigate the dangers of submitting a morphed image into the enrollment stage of the biometric system pipeline [1]. In their work, they created a small dataset of high quality morphed images by hand using the image editor GIMP. They show high attack rate morphed images can be generated using the likeness of two or more look-alike subjects to create an ambiguous face [17].

Morphs pose a serious security problem. If a bad actor is morphed with a look-alike, the synthesized image can be shared between the two people. Morphed images are commonly generated with two separate methodologies, landmark [9, 10, 18] and GAN-based models [11, 19, 12]. Both of these methodologies have their own strengths and weaknesses [12].

One of the major pitfalls for landmark-based morphing is the number of ghosting artifacts in the high-frequency regions. Chaudhary *et al.* [4, 5] showed the possibility of detecting morphed images solely in the wavelet domain by analyzing the high-frequency subbands where artifacts are likely to occur. Using this information, we propound a new method of morphing using the wavelet subbands to blend the morph image in order to

[1] Portions of this section are taken from the prior work of [13]

decrease the likelihood of artifacts in the high frequency regions. In our new morph generation method utilizing the Discrete Wavelet Transform (DWT), the input images are warped and then subsequently blended using their wavelet decomposed sub-bands (see Fig. 2.3). Our work is the first that leverages the spatial-frequency wavelet domain to create high quality morphs [13].

High-quality morph images should be of high perceptual quality and be able to fool FRS. To create morphs of the highest attack rate, we apply adversarial perturbation to the morphed images in order to further increase verification difficulty by fooling a classifier into mislabeling the examples into an erroneous class. The added perturbation should be unperceivable in the image. After the perturbation is applied, the morph will be realistic enough to fool a human while having a high attack rate when presented to a FRS.

2.1 Passport Photo Standards

The greatest national security threat morph images pose is the possibility of a bad actor submitting a morph image as a passport. The U.S. Department of State imposes strict guidelines that are required of an image to meet these standards [20]. For a morph photo to be a realistic threat, it must meet the composition and size requirements of a passport photo. A digital passport photo must meet the following requirements:

- Image must be a square.
- Image size must be no smaller than 600×600 pixels.
- Image can not have evidence of JPG compression.
- Images can not be scaled up to meet size requirements.
- The head must take up between 50-69% of the height of the image.

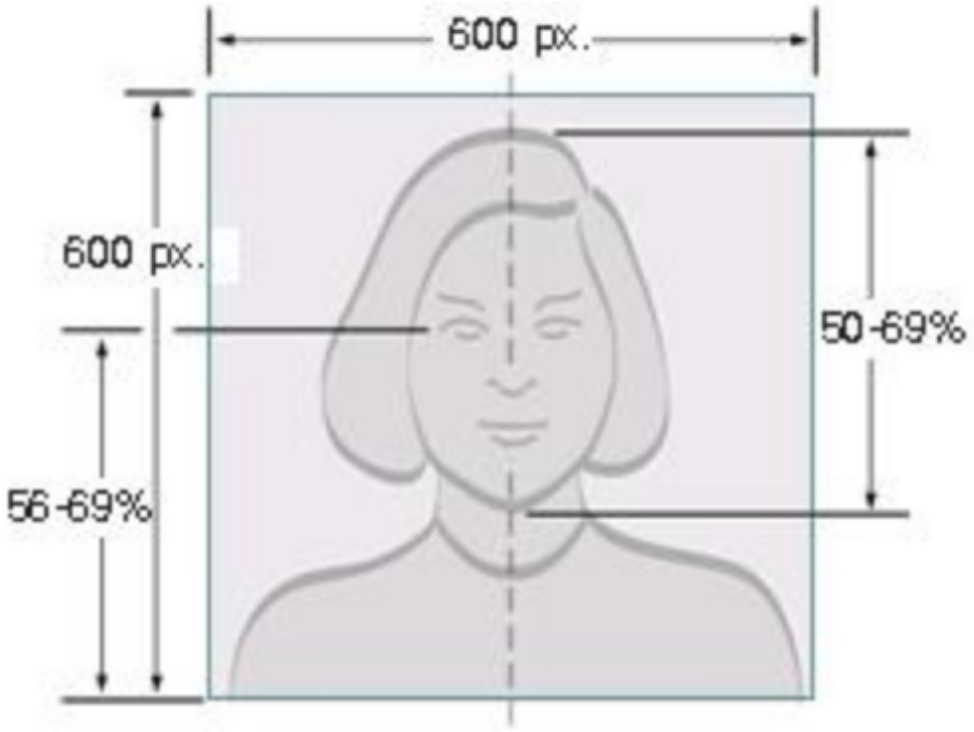

Figure 2.1: Required composition for a digital U.S. passport photo as outlined by the U.S. Department of State.

- Eyes of the face must be between 56-69% of the height of the image.
- Subject can wear eyeglasses.
- Subject must be posed in front of a neutral background.
- Subject must be looking directly into the camera with a neutral expression.

In order to meet these requirements, potential datasets used to make morphs must be properly vetted. Any image of a face must first meet the composition requirements laid out above. For example, the subjects must be looking directly into the camera with a neutral background for them to be considered. This requirement eliminates many popular datasets used in literature that feature people in the wild. Example images of proper vs improper

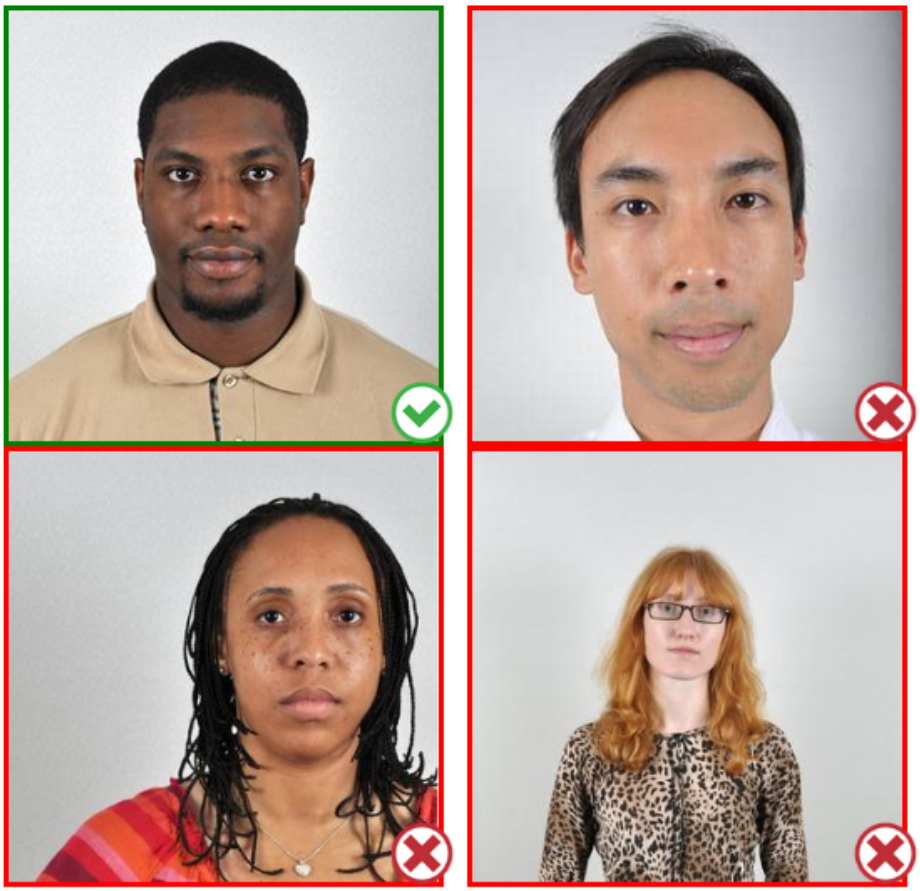

Figure 2.2: Example images of proper passport photos.

passport photos can be seen in Figure 2.1. Next, the face in the photo must meet the size requirements. In many images, it is not possible to properly crop the face of the image without upscaling the photo to the minimum 600×600 pixel size.

Once it is decided that a dataset meets the quality requirements of a passport photo, we run the dataset through a passport cropping tool. We created a custom cropping tool that automatically checks the face's composition in the image, and if the face meets the required size and position requirements, crops the face. All faces that do not pass the requirements of the cropper are discarded. This step is done to every image prior to morphing. Not only does cropping the images create passport-style images, it also ensures that we have a high-quality dataset for morphing. If a face is not looking at the camera or improperly

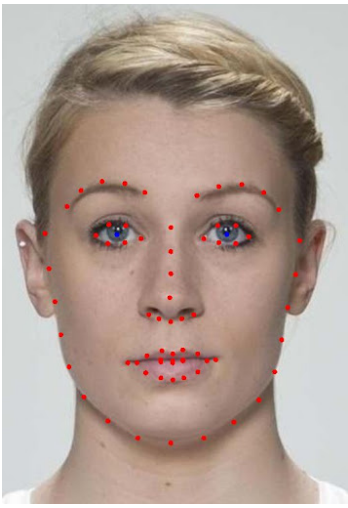

Figure 2.3: Landmark points found on subject.

positioned in the image, it will create morphing artifacts and additional shadowing in the landmark morphs. Further, because all data being used to generate the morphed images already meets passport standards, the resulting morph images will as well.

2.2 Landmark Morphing

The landmark-based morphed image generation typically consists of three steps: landmark detection, warping, and blending. The landmark points of the two input subjects, which are critical points on each face, are averaged together to create common landmarks. The images are then warped towards these common landmarks and blended to create the morphed image. The morphed images are guaranteed to have visual similarity with both individuals because features of the individuals are combined by averaging the input images together. Ferrara [17] morphed their images manually using the open-source image editor GIMP. While the resulting images showed little artifacts, the pipeline was tedious and inconvenient to be scaled up for generation of large datasets. Since then, many open-source repositories have emerged, making it simple to generate large-scale datasets. Sarkar [12]

generate three morphed datasets utilizing four popular morphing repositories: Facemorpher [9], OpenCV [10], WebMorph [18], and StyleGAN2 [14]. Facemorpher, WebMorph, and OpenCV are typical landmark-based algorithms that rely on a combination of warping and splicing to generate morphed images. While landmark-based morphing techniques are fast and effective, they tend to lead to warping artifacts in the high-frequency areas in the image such as around the iris and outline of the face [21]. To have a successful morphed attack, there should be no visible artifacts in the image.

When generating images, we consider two look-alike individuals for morphing. The pair's faces, $\boldsymbol{u}$ and $\boldsymbol{v}$, are aligned. 68-element long pixel-coordinates $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are found on each subject's face. An example of these coordinates can be seen in Figure 2.2. The landmark coordinates are areas deemed of high importance for morphing. Then, $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are used to generate a mesh grid across the image. On an element-wise basis, the coordinates of $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ are averaged together to create the common landmarks coordinate, $\hat{\boldsymbol{m}}$. After warping to the common landmarks, bilinear interpolation is performed in order to correct color values. An affine transform is used to transmute points from $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ to the $\hat{\boldsymbol{m}}$ creating both $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. After warping, $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are averaged together. At this point, the background of the face regions will have a heavy ghosting effect. The face region is spliced from the background and placed onto the convex hull of $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ to generate the final image $\boldsymbol{m}$. Our algorithm is modified from both Facemorpher [9] and OpenCV [10] at the stages where the background is warped and where the convex hull is spliced.

When the warped images are blended together, high frequency areas of the images tend to create a shadowing effect. This occurs when sharp edges in the faces such as around the areas of the faces are not perfectly aligned and create artifacts. Examples of these artifacts can be seen in Figure 2.5. The image on the right is a common artifact seen, where the iris of the two individuals does not warp properly, thus, creating a "second iris". The middle image shows the ghosting effect that occurs when one subject has a mustache

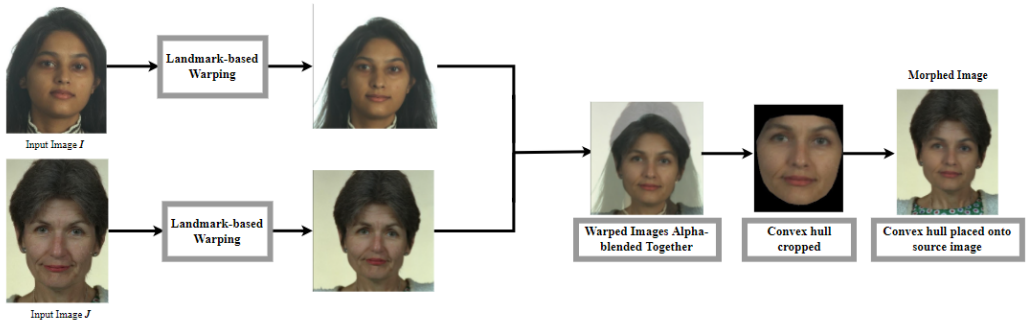

Figure 2.4: Standard landmark morphing pipeline where the two input images are warped toward one another and then blended.

and the second subject does not, creating an unnatural looking tint on the skin. Anomalies like glasses, moustaches, and stray hair can cause these types of artifacts, highlighting the need for a high-quality dataset that has been thoroughly pre-processed for anomalies in subjects. In their work, Sarkar *et al.* selectively pair individuals based on the types of anamolies the subject poses. Lastly, the image on the right shows another common issue where the eyebrows of the faces are not properly aligned, creating a double eyebrow effect. These artifacts make a morphed image easily identifiable. In order to make high-quality morph images, there should be no obvious artifacts in the morph image. In order to minimize artifacts in the generated images, the pair of subjects selected for morphing should naturally look alike. Pair selection is a vital step when morphing faces and can lead to drastic differences in quality of the morphed image [22, 23]. If the morph is to be passed between two individuals, the individuals must possess physical similarities. Scherhag [23] proposed to classify a dataset into soft-biometrics such as hair color, skin color, age, and gender prior to morphing. Damer [22] explored the different methods of determining look-alikes and how they affect the quality of the morphs. They find a strong correlation between morphing similar looking individuals and higher attack rates.

Figure 2.5: Common blending artifacts found when landmark morphing faces that occur when features are not properly aligned.

2.3 Wavelet-based Morph Generation

In order to provide greater flexibility and improve the blending stage of the landmark morphing technique we introduce a new approach to fusing the warped images. Our second method of landmark-based morphing leverages the spatial-frequency decomposition to fuse the warped images. After warping, we decompose the images into 64 wavelet subbands using Discrete Wavelet Transform (DWT).

The look alike pair of images are aligned and warped in the same manner as Section **??**. However, after the warping stage, $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are decomposed into 64 sub-bands using a three-level undecimated wavelet decomposition. Vertical and a horizontal filters are applied to the warped images, creating the Low-Low, Low-High, High-Low, and High-High sub-bands. We number the bands from $1, 2, \ldots, 64$, where the first sub-band represents the baseband. As presented in Fig. 2.3, the lowest frequency baseband after three-level wavelet decomposition of the $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ are averaged together. This sub-band is selected because it represents most of the shared information from the original subjects. The remaining 63 sub-bands are combined using the maximum-coefficient at every location in the sub-bands to capture the most significant information from each subject. If $[\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{64}] = \Phi(\hat{\boldsymbol{u}}_{\boldsymbol{w}})$ and $[\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{64}] = \Phi(\hat{\boldsymbol{v}}_{\boldsymbol{w}})$ are the undecimated wavelet decompositions of the aligned

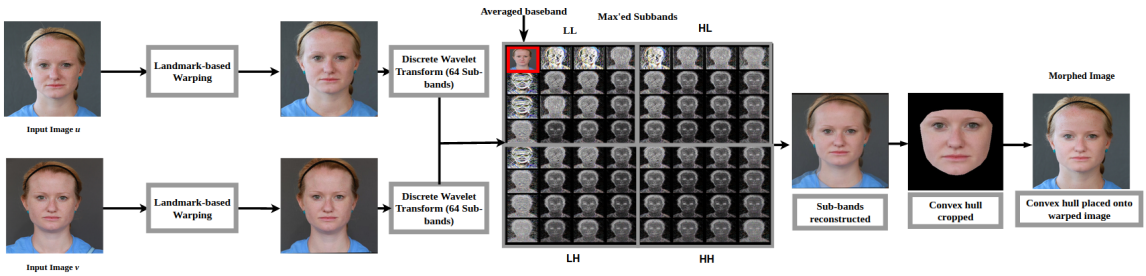

Figure 2.6: Standard landmark morphing pipeline where the two input images are warped toward one another and then blended.

input images, we define morphed sub-bands as:

$$\boldsymbol{M}_k[i,j] = \begin{cases} \text{mean}\left(\boldsymbol{U}_k[i,j], \boldsymbol{V}_k[i.j]\right), & k=1 \\ \max\left(\boldsymbol{U}_k[i,j], \boldsymbol{V}_k[i.j]\right), & \text{otherwise.} \end{cases} \tag{2.1}$$

These morphed wavelet sub-bands are used to reconstruct the blended face image. The convex hull of the morphed image is spliced onto the background of $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ to create the morphed images. It is hypothesised that by maxing the high frequency areas, it will help retain sharp edges from a single subject, in order to help eliminate shadowing and by averaging the low frequency region, color and general shape of the face are retained.

2.4 GAN-based Morph Generation

In addition to landmark-based techniques, we utilize a Generative Adversarial Network (GAN) to produce morphed images. GAN-based morph generation creates morphs by combining the latent space representation of two face images. GANs were first introduced by Goodfellow [24] who introduced the idea of an adversarial network to generate high-quality images by training two multilevel perceptrons simultaneously minimizing the objective function of a generator and maximizing the objective function of a discriminator. In

other words, the generator is trained to synthesize data that fools the discriminator, while the discriminator is trained to detect artifacts generated by the generator [24]. Since then, GANs have made strides in terms of quality and accessibility [19].

Today, one of the most common GANs for morph generation in the literature is StyleGAN2 [14] because of its high-quality results and minimal artifacts. This GAN-based approach uses latent vectors of input images which are then linearly combined, resulting in minimal artifacts and producing high-quality morphs [12, 14, 25]. Damer introduced MorGAN [11] for face morphing. MorGAN utilizes an encoder which is jointly trained with their discriminator and generator in order to learn the mappings between the encoder and decoder. The networks are trained to generate high-quality reconstructions from the bottleneck. Once MorGAN was trained, the latent vectors were linearly combined in order to generate the morph image. GAN-based morphing approaches have issues retaining identity information, causing morphs to be more heavily weighted toward one subject than another [12, 21]. However, a morphed face should pass as either input subjects resulting for an effective morph generation algorithm. MIP-GAN [21] attempts to fix this problem by creating a loss-function based on perceptual-loss and identity priors in order to retain similarity to input subjects while creating high-quality morphs. This methodology was met with success, as the MIPGAN based morphs can fool multiple FRS at a higher rate than StyleGAN-based morphs. Damer *et al.* introduced MorGAN [11] for face morphing. They utilize their discriminator and generator in order to learn the mappings for the encoder and decoder. The networks are trained to generate reconstructions from the information bottleneck. Once MorGAN was trained, the latent vectors were linearly combined to generate the morphed image. We combine the latent code using StyleGAN2 [26] to generate our morphed images because of the high-visual quality of their output images. While GAN-based approaches are becoming more popular, literature shows that GAN-generated morphs struggle to retain the identity of the input subjects [12]. Identity retention is vital

for morphed images because the morphed face should be able to be verified between both input subjects.

The same aligned pairs as described in the previous section are used as $\boldsymbol{u}$ and $\boldsymbol{v}$. They are warped toward common landmarks in the same manner to result in the warped faces $\hat{\boldsymbol{u}}_{\boldsymbol{w}}$ and $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$. The face region of both warped images are spliced and pasted onto a black background. These images are embedded to an 18×512 latent code. These codes are then averaged together to construct the morphed image's latent code. To improve final visual quality of the morphs, custom noise is added to the convolutional layers of StyleGAN2. This fused latent vector is reconstructed to generate the morphed convex hull. This face image is spliced back onto the face region of the input images $\boldsymbol{u}$ to construct the morphed image $\boldsymbol{m}$.

2.5 Adversarial Perturbation

Adversarial perturbation is added to the morph images with the intention of fooling a morph detector into labeling the input as a bona fide class. Typically, the pixel values are constrained to an L_∞ value which help to preserve the quality of the perturbed image. Adversarial perturbation should not be perceptually visible in the final image. Goodfellow et al. [27] introduce the fast gradient sign method (FGSM), which perturbs the input of the model based on the sign of the gradient for a target class. Liao et [28] utilized FGSM with a masking technique to perturb areas deemed as high importance using spatial information derived from multiple convolutional layers in a model. Hussain et al. [29] leverage adversarial perturbation for their work on adversarial deepfakes by perturbing frames of a video labeled as fake by a detector with the intention of all output frames being labeled as real.

We train our model to detect morphed images based on the work from the authors of [30]. The morph detector is able to detect morphs with near perfect accuracy. We add

adversarial perturbation to the morphed images in order to further increase difficulty of their detection. FGSM perturbs an image based on the gradient with every iteration of backpropagation. Basic Iterative Method (BIM) [31] is a derivation of FGSM, where a constant step-size is utilized for every applied perturbation and an L_∞ constraint is used as maximum allowed pixel difference. Using our trained morph detector and BIM, images are perturbed:

$$\boldsymbol{m}_{N+1}^{adv} = \mathit{Clip}_{m,\epsilon}\{\boldsymbol{m}_N^{adv} + \beta \text{sign}\left(\nabla_m L_{adv}\right)\}, \tag{2.2}$$

where $\boldsymbol{m}_0^{adv} = \boldsymbol{m}$ is the morph and L_{adv} consists of cross-entropy and Total Variation (TV) smoothing losses:

$$L_{adv} = J\left(\boldsymbol{m}_N^{adv}, y_{true}\right) - \lambda TV\left(\boldsymbol{m}_N^{adv}\right), \tag{2.3}$$

where J is the cross-entropy cost function between the adversarial image and the target class, β is the perturbation step size and ϵ is the L_∞ constraint on the pixel difference values [31]. The term y_{true} is equal to 1 and 0 for morph and real images, respectively. The value of $\mathit{Clip}_{m,\epsilon}$ confirms that the pixel values are within ϵ L_∞-norm distance from the original sample. We also clip the adversarial example at each iteration to make sure that all pixel values reside within the valid input range. Variable λ is the smoothness regularization parameter. To further improve the visual quality of the image, TV smoothing is applied to the perturbation image to remove any visible artifacts in the adversarial morphed image [32, 33]:

$$TV(\boldsymbol{m}_N^{adv}) = \sum_{i,j}((\boldsymbol{r}_N[i,j] - \boldsymbol{r}_N[i+1,j])^2 - (\boldsymbol{r}_N[i,j] - \boldsymbol{r}_N[i,j+1])^2)^{\frac{1}{2}}, \tag{2.4}$$

where $\boldsymbol{r}_N[i,j]$ is a pixel in the perturbation image $\boldsymbol{r}_N = \boldsymbol{m}_N^{adv} - \boldsymbol{m}$. We refer to the perturbed morph image as $\boldsymbol{m'}$.

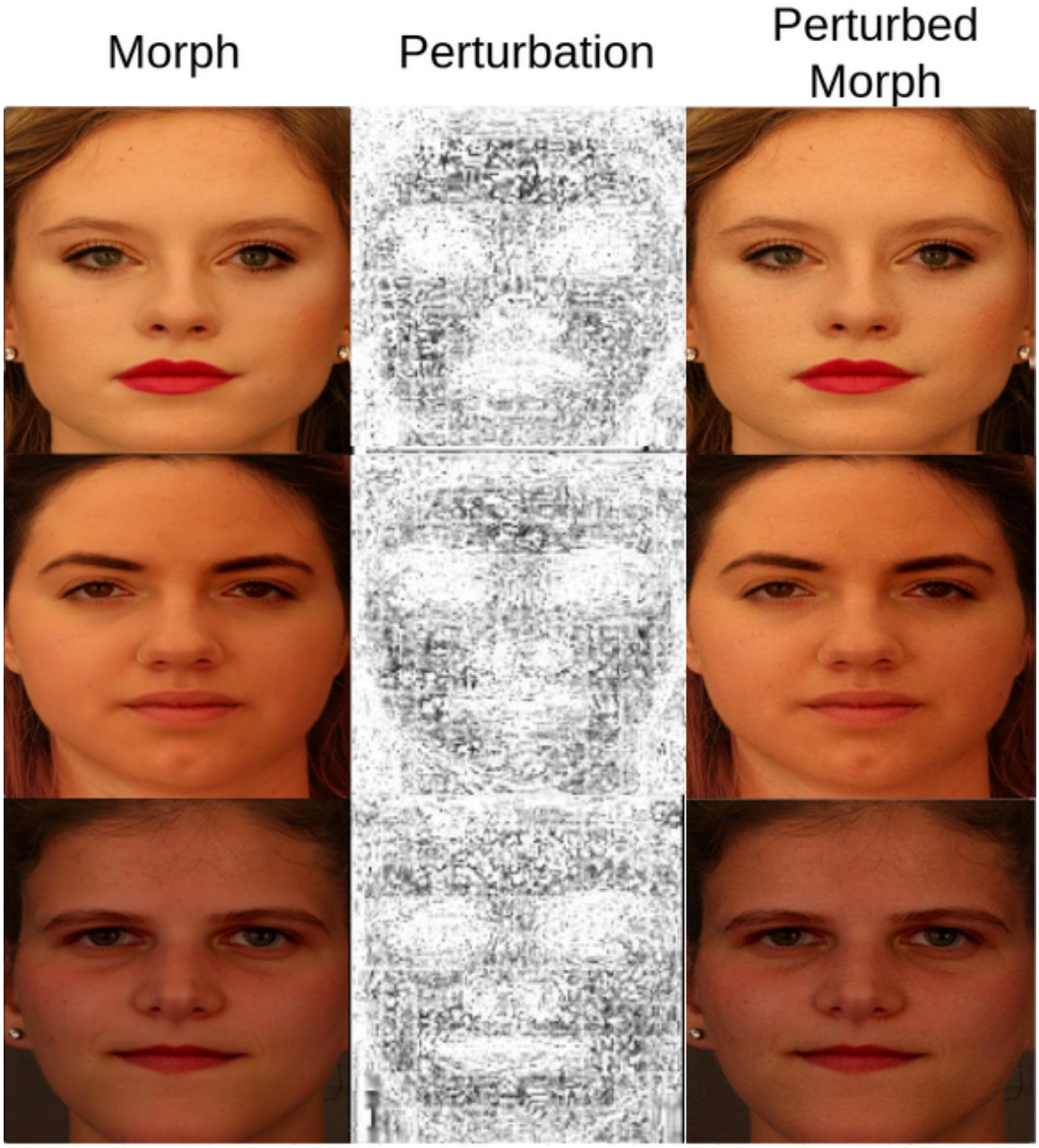

Figure 2.7: Examples of perturbed images.

2.6 Morph Detection

There are two main methodologies of morph detection, single and differential. Single morph detection is effective at determining morph artifacts such as shadowing or artifacts described in Section 2.2. Differential morph detection detects morphs in a verification scenario and represents the quality of morph in terms of identity between the bona fide subjects and the morph image. In order for the morph image to be effective, it must be free of both artifacts and look like both contributing bona fide identities. A major barrier for morph detection research is the lack of publicly available high-quality morph datasets.

There are two trains of thought when using classical approaches to morph detection. The first utilizes methodologies introduced using image forensic techniques to detect tampering in the image. The second train of thought focuses on texture analysis. Typically, these classical methods segment the image and then deploy feature descriptors with a Support Vector Machine (SVM) to classify images [4, 5, 34]. Common feature descriptors include Speeded-Up Robust Features (SURF) [35], Local Binary Patterns (LBP) [36], Scale-Invarient Feature Transform (SIFT) [37], Binarized Statistical Image Features (BSIF) [38], and Histogram of Gradients (HOG) [39].

Classical methods of morph detection have been nearly entirely replaced by deep learning-based algorithms. Chaudhary *et al.* showed promising results when detecting morph images decomposed to the wavelet domain in a differential setting [4]. Their results have EERs of significantly lower than SURF, SIFT, LBP, BSIF, and FaceNet across four datasets found in literature. Scherhag *et al.* [23] introduces a hybrid approach to morph detection. They utilize both classical (texture descriptors, keypoint extractors, gradient estimators) and a deep learning-based network to extract features, which are sent SVMs and then fused together to classify the image.

Chapter 3

Experiments

Three datasets were generated and testing, the Baseline Wavelet dataset, Child dataset, and Twin dataset. First, we created our Baseline dataset using the FRGC [16] and FERET [15] datasets. These datasets are compared to the morph images found in literature using the same datasets generated by Sarkar *et al.*[12]. We show the effectiveness of our wavelet morph generation as well as the effects of adversarial perturbation applied to morph images. Next, in order to create the most difficult morph images possible, we morph identical Twins. The advantage of morphing twins is two fold, 1) Pairing twins removes ambiguity in morph pair selection, 2) Identical twins already look alike, leading to less warping compared to other dataset and thus creating less artifacts. Last, we generate and analyze the Child dataset. The child morphs take advantage of the fact that FRS are not properly trained on children, and by generating morphed children, it takes advantage of this gap in algorithm training. All datasets meet this ICAO standards for passport images [3].

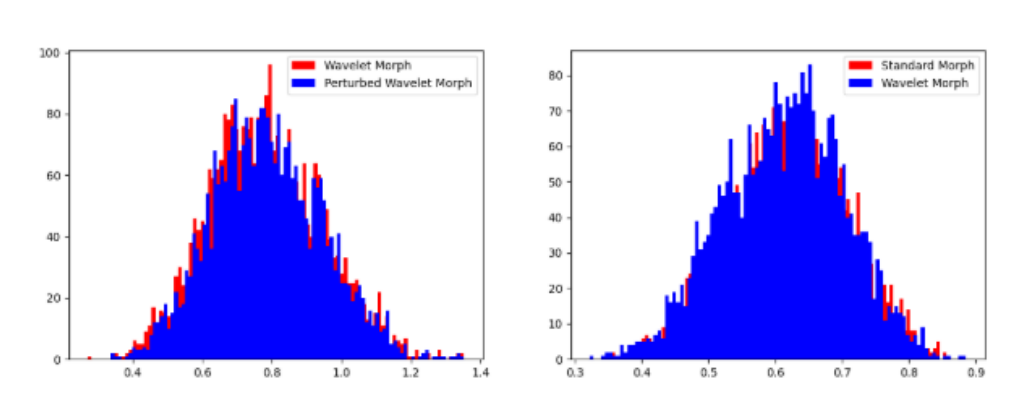

Figure 3.1: FaceNet L_2 distribution (left) and SSIM comparison (right) between input subjects and their respective morphs.

3.1 Wavelet Morph Generation

The first morphed face images were created in order to compare to morphing datasets to those found in literature [12] [1]. We used an adapted version of the Facemorpher [9] methodology to morph our images to create wavelet-based morphed images as described in [13]. In order to further improve the morphing capability of the morphed images, we adversarially perturb the images using a morph classifier. We compare our work to the work done by Sarkar [12], where they generated a Facemorpher generated dataset from the FERET [15], FRLL [40], and FRGC datasets [16].

We utilize the FERET, FRLL, and FRGC datasets for our morphs [15, 40, 16]. The datasets contain image sizes of 413×531 for FRLL, 1704×2272 for FRGC, and 512×768 for FERET. Each dataset depicts passport style images with a neutral face looking into the camera under ideal lighting conditions. In total, FERET contains 1,199 different identities, FRLL contains 102 identities, and we used a subset of FRGC which contains 765 identities. We use morphed images from [12] for comparison to our generated morphs. We refer to these image as the standard morphs for the rest of this paper because they use

[1] Portions of this section are taken from [13].

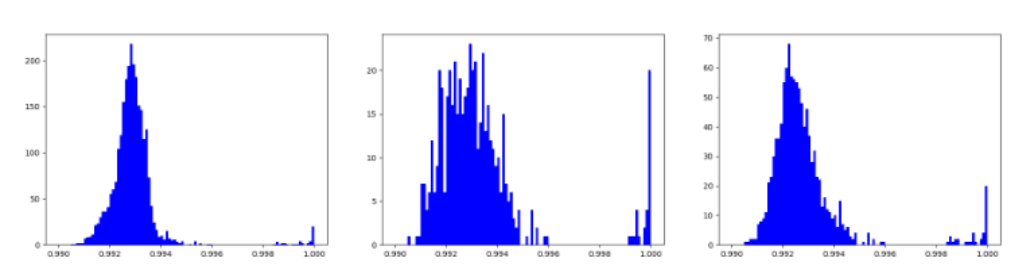

Figure 3.2: SSIM distributions between wavelet morphs and their respective perturbed wavelet morphs for (left) FRLL, (middle) FERET, and (left) FRGC datasets.

a typical morphing pipeline consisting of an alpha-blending step used to combine warped images. For pairing our morphs, we use the protocols originally created by Neubert *et al.*'s AMSL dataset for FRLL [40] and Scherhag *et al.*'s protocol for FERET and FRGC [34]. In addition, landmarked-based wavelet morphing images described in Section III.A are referred to as wavelet morphs.

3.1.1 Similarity Comparison

Morphed images share features from both input subjects; therefore, a quantitative measure of perceived similarity is needed for comparison. We use two different metrics, a FaceNet match score [41] and the Structural Similarity metric (SSIM) [42]. Both metrics are selected because they represent a perceived similarity rather than a direct pixel-comparison to their bona fide subjects. FaceNet is leveraged to quantify look-alikes as deep learning-based verifier, while SSIM uses classical techniques to quantify perceptual similarity. To minimize extraneous information from the morphed images, the convex hull region of the morph is extracted for the comparison and placed on a back background.

FaceNet uses a deep convolutional network architecture to create a compact feature embedding of its input. FaceNet is trained using triplet loss, where the Euclidean distance (L_2) for embeddings of the same identity are positive examples and differing identities

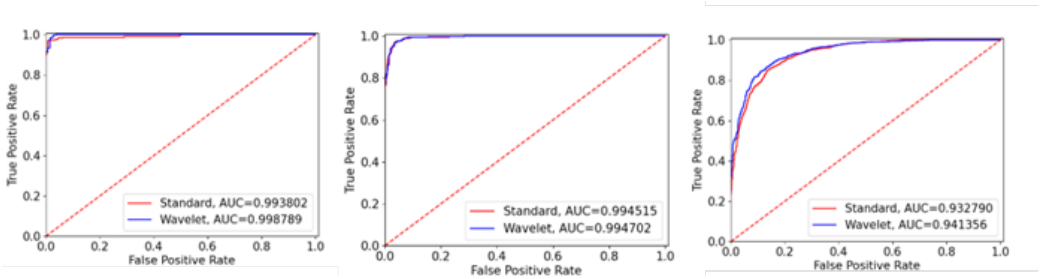

Figure 3.3: Differential morph detection: ROC curves for the (left) FRLL, (middle) FERET, and (right) FRGC datasets.

are considered negative examples [41]. Therefore, there is a correlation between the L_2 distance of feature embeddings and perceived similarity. SSIM is a quality metric used to mimic similarity of two images as perceived by the human eye. SSIM is calculated using a combination of three independent comparisons: luminance, structure, and contrast [42]. Visible artifacts in an image decreases the SSIM.

Figure 3.1 shows the distributions of the L_2 distance between FaceNet embeddings and SSIM comparison between Sarkar's standard landmark morphs and our wavelet morphing technique. For both distributions, every morph has two separate comparison values, for each contributing bona fide identity. For our FaceNet comparison, a smaller L_2 value represents a stronger look-alike. Inversely, a larger SSIM value represents a stronger look-alike. Both distributions show that our wavelet-based morphing technique is as effective at creating morphs that look like their input subjects as [12], while retaining the image quality. The standard and wavelet-based morphs share the same mean of their SSIM distributions (SSIM of 0.61), resulting in no difference between visual quality of the standard and wavelet-based morphs.

To test the attack effectiveness of our wavelet morphs versus the landmark morphs, we use FaceNet as a verifier to perform differential morph detection. The results from FaceNet are shown in Figure 3.3.

3.1.2 White-box Detector and Verification

As presented in Figure 3.3, FaceNet can differentiate between morph and genuine images when compared to a reference photo for both our wavelet morph and Sarkar *et al.*'s morphs [12] . The wavelet-based perturbed and standard images are tested on the trained morph detector and the results are shown in Figure 3.4. In Equation 2.2, we use $\beta = 6$ and $\epsilon = 2$ for perturbation. Image perturbations take approximately 2 seconds per image. The AUC for FRGC is 67%, FERET is 24%, and FRLL is 2%. The results show that the perturbed images are being erroneously classified as bona fide images at an alarming rate. Figure 3.2 shows the distribution of SSIM values. If the images are too heavily perturbed, they exhibit signs of degradation. The SSIM score for all datasets show that every perturbed image has an SSIM of above 0.99, indicating that all images are perceived to be indistinguishable to the wavelet-based morph. Our perturbed wavelet morphs would bypass a morph detector in the passport pipeline with a high degree of success.

To determine the morph's effectiveness on the verification stage, we utilize a pretrained FaceNet model as a verifier [41]. A reference image of a subject is compared to a second genuine image of the subject to create a positive comparison, and a negative comparison is made between the reference image and its respective morph. FaceNet ROC curves are plotted for each of the datasets as shown in Figure 3.4. The true positive score signifies a morph correctly labeled as a morph. FaceNet discerns the wavelet morphs at a nearly identical rate as the standard morphs. The verification stage is the most likely point in the pipeline for the morph to be detected. Verification is a difficult problem for face morphing because the morph image must contain features from both input subjects, making it difficult for the resulting morph to appear more similar to a reference image than to a bona fide image.

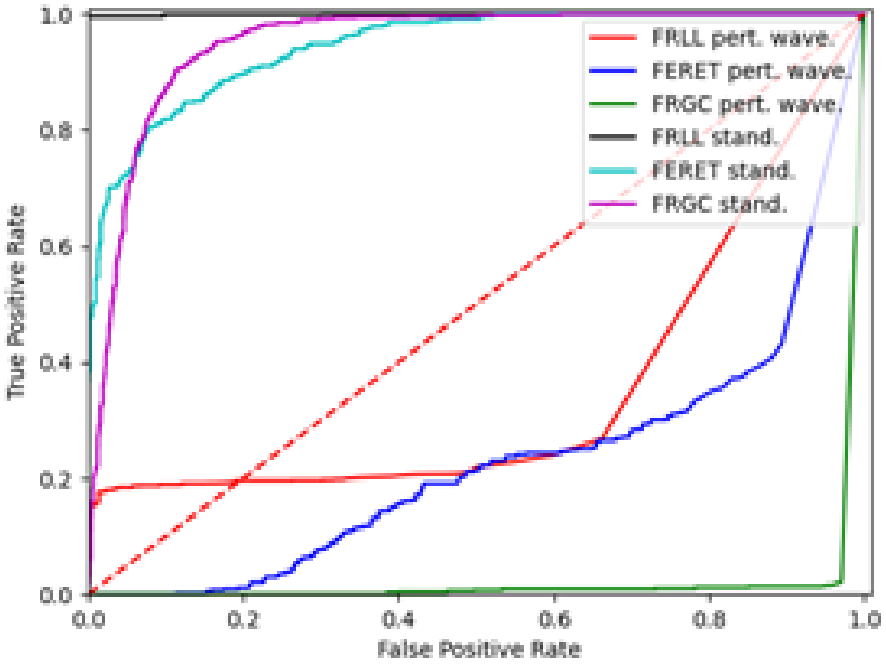

Figure 3.4: Classification of our morph datasets compared to those used in literature.

3.1.3 Summary

In this section, we provided the prospect of morphing in the spatial frequency wavelet domain. We showed that our wavelet-based morphs are as convincing as morphs generated in prior works, while introducing a new morphing methodology. Wavelet selection offers additional flexibility when blending images that was not possible before. By adding adversarial perturbation, the wavelet-based morphs are nearly impossible to detect by humans and deep learning-based detectors. In the future, more sophisticated methods of sub-band selection can be used to generate morphs in the spatial-frequency domain, creating morphs that are more difficult to detect. Further, more work must be done to better understand the transferability of the perturbed images.

3.2 Twin Morph Generation

Identical twins, also known as monozygotic twins, pose a severe problem to FRS since they represent the extreme scenario between individuals who naturally look alike [43]. Finding look-alikes is a necessary step when creating high-quality morphs [22] in order to reduce artifacts and improve verification properties. Paone *et al.* [43] studied a twins dataset made up of 126 twin pairs. They found that the Equal Error Rate (EER) for a twins dataset is significantly high. Five of the seven algorithms tested had an EER at or above 50% for identical twins. Therefore, identical twins are a challenging paradigm for an FRS because of twins' inter class similarity which can lead to high false acceptance rates in the verification stage [4, 43].

Identical twins represent the ideal pairing condition for morphing and remove the ambiguity of pairing look-alikes. The effectiveness of morphing twins is two-fold. First, Commercial Off-The-Shelf systems (COTS), as well as human verifiers, are vulnerable to the high-quality morphing attacks generated from similar face images [44]. Second, twins naturally looking similar creates ambiguity between individuals, causing an increase of false acceptance in detectors [4] and creating a very useful dataset for training and testing morph detectors.

To create an extremely hard scenario for an FRS, we generate a new dataset of identical twin morphed images. Our morphed faces are generated with three separate methodologies, landmark-based models [9, 10], Generative Adversarial Network (GAN)-based models [45, 14], and our wavelet-based morphing. Our Twin dataset provides ideal look-alike pairs for morphing. Consequently, we observe that the Twin dataset provides better morph generation capability compared to several other datasets across different morphing methodologies [12, 14]. As shown in Figure 3.5, for the same FaceNet [46] distance between the morph and bona fide, the twin morph dataset looks significantly more similar to its con-

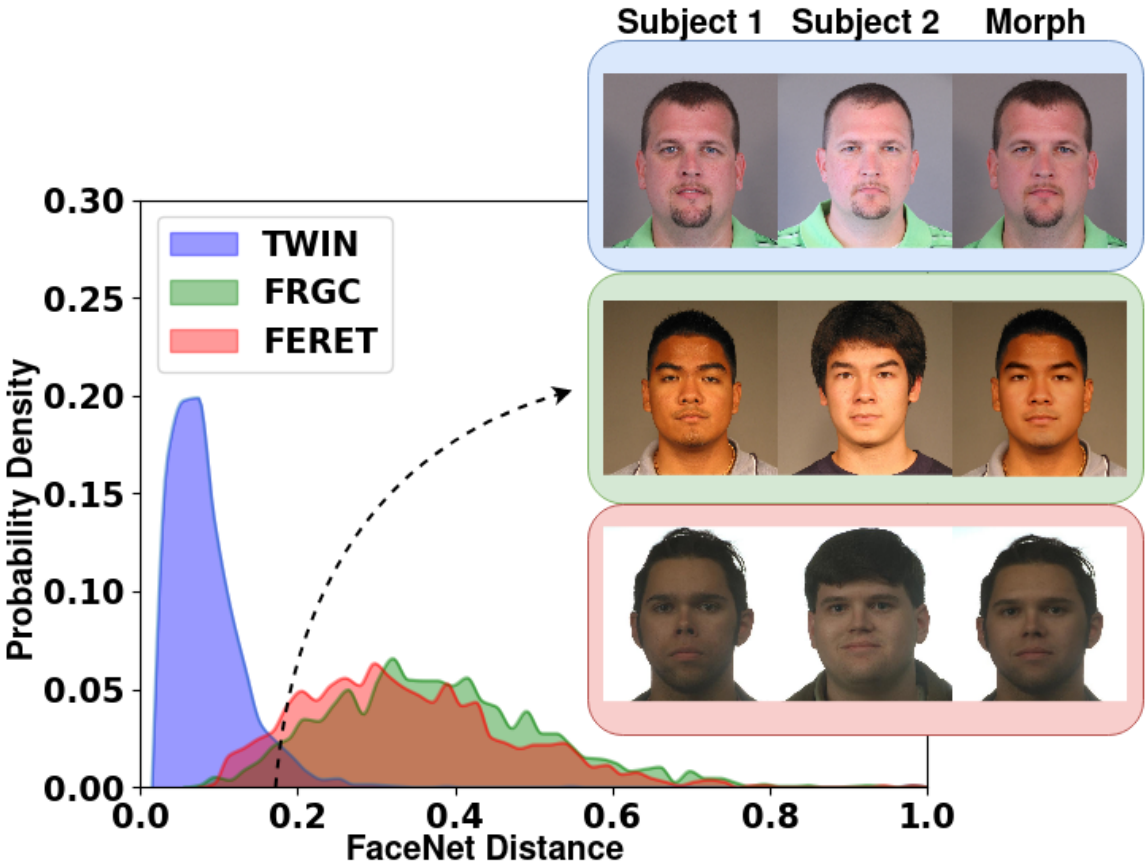

Figure 3.5: Probability density of the normalized FaceNet L_2 distances between the embeddings of the morph and their respective bona fide subjects for the Twin, FRGC, and FERET datasets. Facemorpher morph examples shown are a distance of 0.19 to both contributing bona fide subjects for the (top) Twins, (middle) FRGC, and (bottom) FERET datasets.

tributing bona fide identities than comparable morph datasets. Finally, we apply adversarial perturbation to these twin morphed images in order to further increase verification difficulty by fooling a classifier into mislabeling the morphed face images as bona fides.

Differentiating twins is a hard problem for facial recognition systems due to the high similarity between the two subjects [47]. In fact, even humans have a difficult time discerning between twin pairs. Biswas experimented with participants differentiating between twin pairs and images of the same person [48]. They discovered that humans are only able to classify twin pairs versus images of the same individual at an average rate of 78.82%.

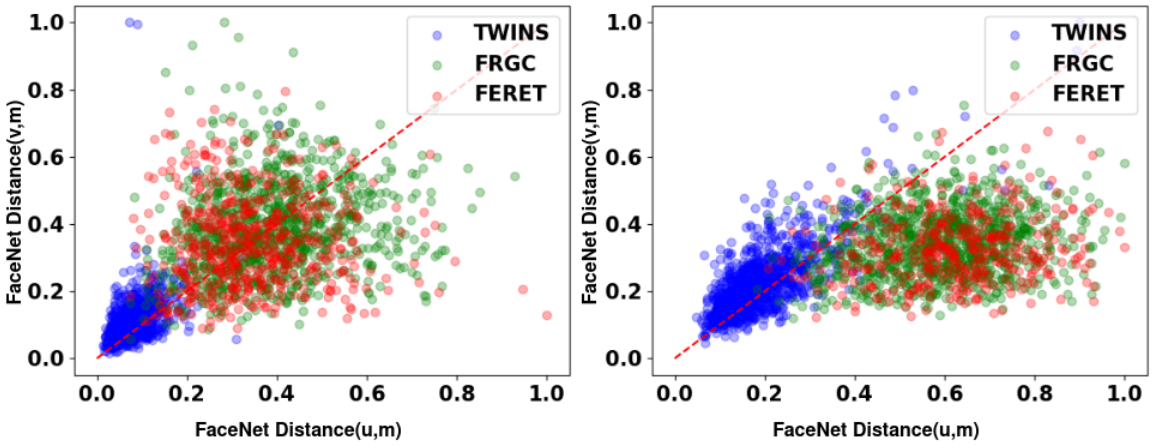

Figure 3.6: FaceNet L_2 distances between the bona fide faces and their respective morphs for the Twins, FRGC, and FERET datasets using (left) landmark and (right) StyleGAN2 morphing methods.

There has been limited research on the effects of twins and facial recognition systems. Paone studied a twins dataset made up of 126 twin pairs. They found that the Equal Error Rate (EER) for a twins dataset is significantly high [43]. Five of the seven algorithms tested had an EER at or above 50% for identical twins.

Pair selection is a vital step when morphing faces, and can lead to drastic differences in quality of the morphed images [22, 23]. If the morph is to be passed between two individuals, the individuals must possess physical similarities. Scherhag [23] proposed to classify a dataset into soft-biometrics such as hair color, skin color, age, and gender prior to morphing. Damer [22] explored different methods of determining look-alikes and how they affect the quality of the morphs. They find a strong correlation between morphing similar looking individuals and higher attack rates. Morphing twin pairs represents the ideal scenario for morphing by removing the ambiguity of pairing look-alikes and guaranteeing high similarity between bona fide subjects.

3.2.1 Vulnerability Analysis

To test our datasets, we utilize the International Organization for Standardization (ISO) [49] standards for reporting the performance, Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER). The ISO describes APCER as the number of morphs incorrectly classified as bona fide presentations in a specific scenario. Inversely, BPCER is described as bona fide images incorrectly classified as presentation attacks [49]. We report APCER and BPCER at the 1%, 5%, and 10%. In addition, we report the Area Under the Curve (AUC) and EER. We also use the Mated Morph Presentation Match Rate (MMPMR) as a metric to quantify the similarity between a generated morph image and its contributing subjects [23] where only morph/bona fide pairs which have a similarity score above a given threshold are considered:

$$\text{MMPMR}(\tau) = \frac{1}{M}\sum_{m=1}^{M}\left\{\left[\min_{n=1,..,N_m} S_m^n\right] > \tau\right\}, \tag{3.1}$$

where M is the total number of morphs and N_m is the number of subjects contributing to a particular morph [23]. S_m^n is the similarity score between the morph m and the nth corresponding subject and τ is the operational verification threshold [23].

The ProdAvg-MMPMR methodology represents the proportion of accepted attempts per contributing subject:

$$\text{ProdAvg-MMPMR}(\tau) = \frac{1}{M}\sum_{m=1}^{M}\left[\prod_{n=1}^{N_m}\left(\frac{1}{I_m^n}\sum_{i=1}^{I_m^n}\left\{S_m^{n,i} > \tau\right\}\right)\right], \tag{3.2}$$

where and I_m^n is the number of samples of subject n within morph m and $S_m^{n,i}$ is the similarity score between the morph m and the ith sample of its nth corresponding subject.

We use FaceNet [46] and ArcFace [52] as our verifiers and the operational threshold is set at False Match Rate (FMR) of 0.1% [53]. MMPMR values for our three morphed datasets are reported in Table 3.10. To calculate the MMPMR(τ) values in Eq. 3.1, for

Table 3.1: MMPMR (%) and ProdAvg-MMPMR(%) for twin morphed datasets.

Method		Landmark	Wavelet	StyleGAN2
MMPMR(%)	FaceNet	97.70	98.11	93.01
	ArcFace	99.20	99.41	94.54
MMPMR-ProdAvg(%)	FaceNet	97.64	97.95	92.97
	ArcFace	99.18	99.20	94.43

Table 3.2: MMPMR (%) at false match rate of 0.1%.

Dataset	Twin			FRGC [12]			
	Facemorpher	Wavelet	StyleGAN2	Facemorpher	OpenCV	StyleGAN2	MIPGAN-II
FaceNet	**97.70**	**98.11**	**93.01**	5.7	5.9	0.7	92.15
ArcFace	**99.20**	**99.41**	**94.54**	11.2	10.8	0.4	94.21

Dataset	FERET [12]			AMSL [50]		LMA-DRD [51]	
	Facemorpher	OpenCV	StyleGAN2	Facemorpher	StyleGAN2	Digital	Print+Scan
FaceNet	40.3	40.6	1.3	81.16	61.28	64.12	60.76
ArcFace	34.8	35.2	2.5	84.85	39.17	80.07	77.17

each subject, we randomly chose a face image that is not used to create the morph. In Eq. 3.2, ProdAvg-MMPMR(τ) is calculated using all the samples corresponding to each subject. The twin dataset includes an average number of 3.15 images per subject. We observe that FaceNet consistently provides lower vulnerability compared to ArcFace and the proposed wavelet-based morphing method provides comparable vulnerability with the landmark morphing method and higher vulnerability than StyleGAN2 model.

3.2.2 Dataset Comparison

For an initial observation of the quality of our morphs, we plot the L_2 FaceNet distance between both bona fide subjects and their respective morph faces found in Figure 3.5. We use the FaceNet distances between the morph and its respective bona fide subjects for the Twin, FERET and FRGC morph[12] datasets. It is clear that the twin morphs have the

obvious advantage when morphing against other datasets, with nearly the entire distribution falling below the spread of both the FRGC and FERET datasets, meaning that the morphs have high similarity to their bona fide identities.

To further understand the relationship between the morphs and their respective bona fide subjects, we plot the same values with respect to both subject 1 and subject 2 in Figure 3.6. The x-axis represents the FaceNet L_2 distance between subject 1 and the morph, where the y-axis represents subject 2 to the morph. We compare the FaceNet distances of the Twins database to the FRGC and FERET morphs. Again, we observe that the twin morphs consistently having lower FaceNet distances than the FRGC and FERET datasets. Further, we observe that the twin dataset has a lower variance in distance scores for both the landmark and StyleGAN2 settings. Meaning that not only do the twin morphs show high similarity, they retain identity significantly better than the FRGC and FERET morphs. This anomaly is especially apparent in the StyleGAN2 datasets, where the FRGC and FERET datasets clearly bias toward one subject.

For the landmark-based datasets in a differential morph detection setting (Table 3.3 and Table 3.4), our twin morphs preform significantly better than the comparison datasets, with our Twin Landmark morphs achieving an EER of 36.84% and 34.36% on FaceNet and ArcFace, respectively. When comparing to the rest of the datasets, all datasets have EER values below 21% across both FaceNet and ArcFace. Using FaceNet, five of the seven datasets (FRGC Facemorpher, FERET OpenCV, AMSL Facemorpher, FERET Facemorpher) have EER values below 5%. Additionally, both the Twin Landmark and Twin Wavelet morphs have an AUC of approximately 14% lower than the best performing landmark comparison dataset using FaceNet (LMA-DRD Digital with AUC of 88.00%). In a differential scenario, it is clear that the twin morphs retain the identity of the bona fide subjects better than the compared datasets.

Table 3.3: Differential morph detection across datasets using FaceNet.

Dataset	Morph Type	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
FRGC Facemorpher [12]	Landmark	99.82%	2.11%	0.63%	0.21%	6.55%	1.22%	0.616%	1.63%
FERET OpenCV [12]		99.00%	18.56%	4.92%	3.03%	14.77%	4.166%	4.16%	2.27%
FRGC OpenCV [12]		99.52%	8.40%	1.60%	0.8%	4.56%	1.52%	0.43%	2.61%
AMSL Facemorpher [50]		99.35%	18.41%	3.70%	1.01%	6.81%	3.31%	1.28%	3.68%
FERET Facemorpher [12]		98.91%	19.45%	5.83%	4.28%	16.97%	4.79%	1.10%	4.79%
LMA-DRD Print+Scan [51]		91.22%	75.68%	33.33%	29.63%	68.56%	41.86%	39.8%	16.27%
LMA-DRD Digital[51]		88.00%	80.26%	45.00%	27.50%	72.36%	57.50%	50.00%	20.00%
Twin Wavelet		**74.03%**	**61.48%**	**57.41%**	**51.51%**	**98.744%**	**85.33%**	**74.56%**	**33.71%**
Twin Landmark		**70.19%**	**64.73%**	**58.31%**	**54.09%**	**99.10%**	**90.20%**	**81.95%**	**36.84%**
FRGC StyleGAN2 [12]	GAN	99.65%	0.40%	0.12%	0.00%	0.42%	0.00%	0.00%	0.42%
AMSL StyleGAN2 [50]		99.98%	1.26%	0.00%	0.00%	0.86%	0.00%	0.00%	0.50%
MIPGAN-II [21]		99.85%	1.89%	0.27%	0.00%	5.06%	0.80%	0.26%	2.40%
FERET StyleGAN2 [12]		99.73%	4.29%	1.56%	0.00%	5.14%	1.56%	0.00%	2.94%
Twin StyleGAN2		**88.92%**	**41.83%**	**33.57%**	**27.77%**	**97.09%**	**62.25%**	**57.62%**	**19.95%**

Table 3.4: Differential morph detection across datasets using ArcFace.

Dataset	Morph Type	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
AMSL Facemorpher [9]	Landmark	97.87%	16.25%	10.78%	6.29%	30.52%	16.82%	6.12%	8.11%
FRGC OpenCV [9]		96.60%	42.85%	15.81%	9.18%	53.18%	19.89%	9.85%	9.74%
FERET Facemorpher [9]		96.33%	26.46%	14.55%	10.77%	57.46%	26.55%	12.66%	10.68%
FRGC Facemorpher [**?**]		96.85%	27.55%	15.85%	11.22%	53.06%	17.34%	10.87%	10.71%
FERET OpenCV [12]		96.32%	24.16%	14.93%	10.77%	57.18%	27.88%	12.47%	10.77%
LMA-DRD Print+Scan [51]		90.26%	67.78%	39.54%	31.97%	53.48%	43.64%	30.32%	17.35%
LMA-DRD Digital[51]		88.88%	65.25%	40.75%	29.63%	57.12%	39.31%	35.70%	21.25%
Twin Landmark		**71.01%**	**81.36%**	**67.44%**	**57.76%**	**98.44%**	**91.77%**	**84.77%**	**34.36%**
Twin Wavelet		**69.98%**	**84.56%**	**71.61%**	**59.93%**	**99.52%**	**92.38%**	**85.27%**	**34.49%**
AMSL StyleGAN2 [50]	GAN	99.96%	0.07%	0.00%	0.00%	0.00%	0.00%	0.00%	0.97%
FRGC StyleGAN2 [12]		99.85%	0.18%	0.06%	0.00%	0.00%	0.00%	0.00%	1.03%
FERET StyleGAN2 [12]		99.75%	3.59%	0.95%	0.37%	10.20%	1.03%	0.05%	2.55%
MIPGAN-II [21]		99.07%	10.65%	6.35%	2.18%	20.25%	8.07%	1.91%	5.91%
Twin StyleGAN2		**93.60%**	**28.39%**	**19.90%**	**15.38%**	**89.67%**	**45.67%**	**22.33%**	**13.58%**

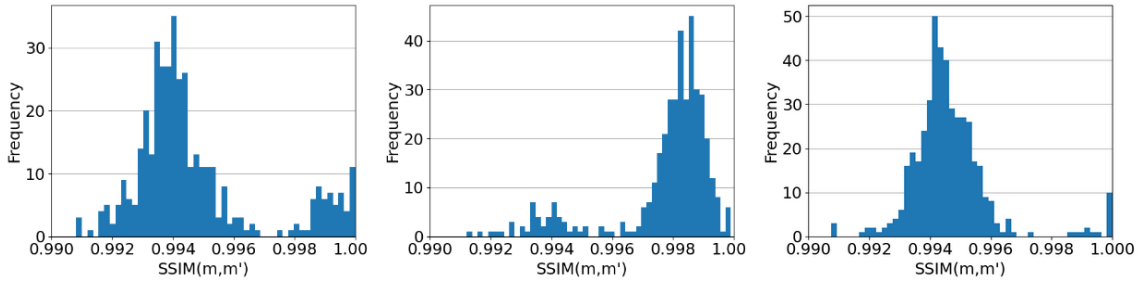

Figure 3.7: SSIM between original and perturbed images for the landmark, wavelet, and StyleGAN2 datasets from left to right, respectively.

3.2.3 White Box Adversarial Attack

Using the architecture of Inception-Resnet v1 [54] pretrained on VGGFace2 [55], we fine-tuned four morph detectors. Our universal detector is trained on all three morph datasets, while the dedicated detectors are trained on each of our three morph datasets to detect morph imagery. We refer to these models as the universal and dedicated trained morph detectors, respectively. We use these detectors to perturb the morph datasets, creating two separate perturbed datasets called the universal and dedicated perturbed datasets. Using a $\beta = 6$, $\epsilon = 2$, and $\lambda = 0.55$, we perturb every morph image until the confidence score of the detector falls below 50%. Prior to perturbation, the landmark, StyleGAN2, and landmark wavelet datasets have a classification AUC values of 95.96%, 99.83%, and 99.80%, respectively. After perturbation, the AUC value of all datasets drops significantly to 46.98%, 56.84%, and 27.87%, respectively for the dedicated datasets. The universal detector was used to perturb the datasets as well, seeing AUC values of 99.30% perturbed to 80.53% for StyleGAN2, 97.90% to 57.73% for landmark, and 99.44% to 55.06% for landmark wavelet to create the universal perturbed dataset. As presented in Fig. 3.8, the perturbed morphed images maintain their visual quality. All the perturbed morphs have a Structural Similarity Index Measure (SSIM) above 0.99 with their morph counterparts which illustrates that the perturbation applied to the morphs is imperceivable.

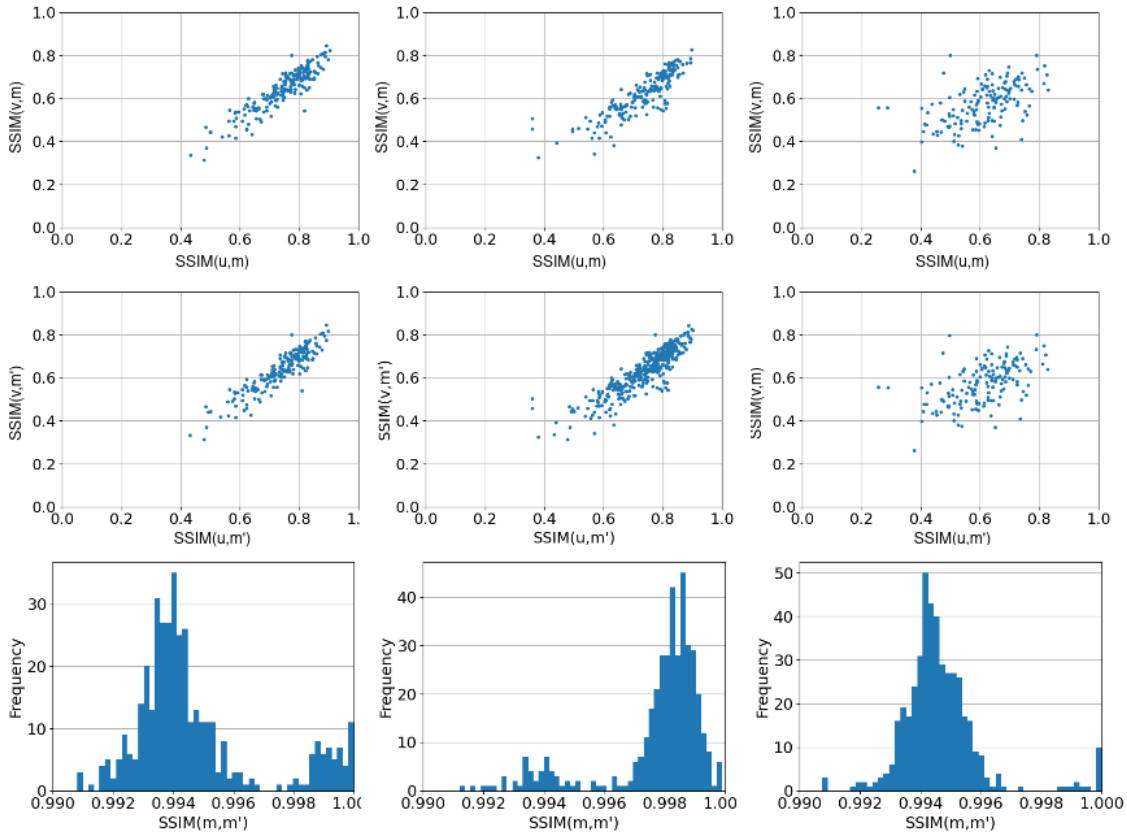

Figure 3.8: The columns from left to right represent landmark, wavelet, and StyleGAN datasets. The top row represents the SSIM score between the 150 randomly selected bona fide and respective morphs. The second row represents the SSIM score between the bona fide and perturbed morphs. The last row represents the SSIM score distribution between the original and perturbed morphs.

3.2.4 Similarity Comparison:

Morphed images contain structural similarities with their bona fide subjects. Our metric of comparison is the Structural Similarity Index Measure (SSIM) which is based on perceived similarity rather than a pixel-to-pixel comparison. This measure is an image quality metric that is calculated by finding the similarity of contrast, luminance, and structure of an image to a reference picture [42]. We compare the SSIM score between the bona fide identities and their respective morphs, as presented in the first row of Fig. 3.8. The land-

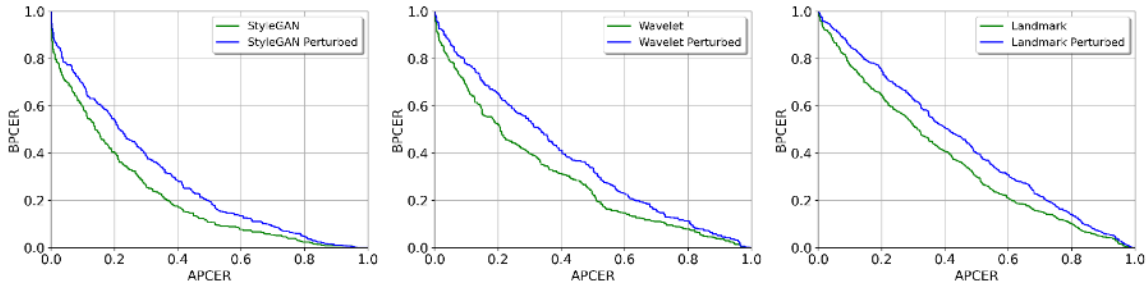

Figure 3.9: Universal FaceNet classifier DET curves for (left) landmark, (middle) wavelet landmark, and the (right) StyleGAN datasets.

Table 3.5: Universal FaceNet classifier tested APCER and BPCER values for morph and perturbed morph images.

Dataset	**AUC**	**APCER@BPCER**			**BPCER@APCER**			**EER**
		1%	**5%**	**10%**	**1%**	**5%**	**10%**	
Landmark	63.62	96.86	90.93	80.81	93.38	86.61	76.69	40.56
Landmark Perturbed	56.22	98.48	93.83	85.23	95.80	92.17	84.67	45.56
StyleGAN2	78.79	95.23	74.18	53.13	79.77	69.55	58.22	28.66
StyleGAN2 Perturbed	71.88	94.65	80.46	69.65	86.66	77.33	68.22	34.26
Landmark Wavelet	70.55	96.61	87.56	75.12	87.50	78.67	67.40	34.55
Landmark Wavelet Perturbed	62.75	97.94	91.78	81.15	92.57	85.56	77.52	39.79

mark and landmark wavelet datasets show a clear, linear correlation between the structural similarities of bona fide identities and the morphed image. While our StyleGAN dataset has a larger variance, the results still show that the identities have a high degree of similarity between both identities. These results reinforce the known issue of identity loss when morphing with StyleGAN [12]. We repeat this test with the bona fide images compared to the perturbed landmark, landmark wavelet, and StyleGAN datasets, as presented in the second row of Fig. 3.8. There was no significant change in SSIM after perturbation and the perturbed morphs retain their similarity with the bona fide identities. Finally, we compare the morphs to their perturbed counterpart in terms of SSIM, as presented on the last row in Fig. 3.8. All the morphs have an SSIM above 0.99 with their perturbed counterparts. This

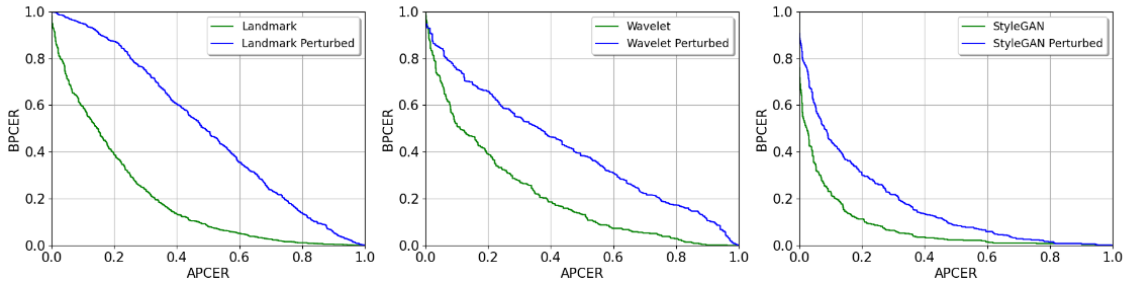

Figure 3.10: Dedicated FaceNet classifiers DET curves for original morph and perturbed morph datasets. (Left) landmark, (middle) landmark wavelet, and (right) StyleGAN datasets.

Table 3.6: Dedicated FaceNet classifiers tested APCER and BPCER values for morph and perturbed morph images.

Dataset	**AUC**	**APCER@BPCER**			**BPCER@APCER**			**EER**
		1%	**5%**	**10%**	**1%**	**5%**	**10%**	
Landmark	80.28	90.00	62.55	47.79	86.12	70.48	58.46	26.53
Landmark Perturbed	50.41	98.37	91.27	85.00	99.43	97.17	94.67	49.35
StyleGAN2	92.20	90.00	36.27	21.86	53.55	32.00	21.11	14.66
StyleGAN2 Perturbed	82.77	94.53	65.11	48.02	78.22	57.33	44.00	25.77
Landmark Wavelet	78.63	88.04	74.63	54.95	87.01	68.38	50.73	28.67
Landmark Wavelet Perturbed	59.10	99.26	95.83	92.64	90.77	84.06	75.00	43.52

illustrates that the perturbation applied to the morphs is imperceivable.

Comparing our StyleGAN2 morphs, the Twin dataset performs with an AUC of 88.92% with FaceNet and 93.00% using ArcFace. The Twin StyleGAN2 images result in the highest EER value across both verifiers. In this scenario, we observe the issue GANs have with retaining identity information, with the EER of all GAN-generated morphs significantly lower than their respective landmark datasets. For instance, the Twin Landmark dataset, using FaceNet as a verifier, has an EER value of 33.71% while the StyleGAN2 generated data has an EER value of 19.95%. This pattern can also be observed in the FRGC Facemorpher and FRGC StyleGAN2 dataset, having an EER of 1.63% and 0.42%, respectively. On the twins datasets, ArcFace is better equipped to differentiate between the morphed im-

age and bona fide subject. For instance, the Twin StyleGAN2 morphs result in an AUC of 88.92% with EER of 19.95% using FaceNet, compared to 93.60% with 13.58% EER with ArcFace. In contrast to FaceNet, ArcFace underperformed with the Twin Wavelet dataset compared to the Twin landmark dataset.

3.2.5 Twin Morph Classification

We extract the FaceNet features and train a two-node binary classifier to classify images as genuine or morph. We train the model using a combined training subset of our three datasets. We call this model the universal FaceNet classifier. A testing subset of the universal perturbed datasets is tested on the universal FaceNet classifier and the results are presented in Table 3.5 and Figure 3.9. Comparing the experiments conducted on 1) real and morphed images and 2) real and perturbed morphed images, we see a drop in AUC for all three morphing algorithms tested: landmark dropping from 63.62% to 56.22%, landmark wavelet dropping from 70.55% to 62.75%, and StyleGAN2 dropping from 70.55% to 62.75%. Additionally, we observe an increase in EER and in APCER@BPCER=5% across all datasets. When ACPER@BPCER increases, the morphs are being labeled as bona fide at a higher rate. This shows that the quality of the morphing attack is improved after adding perturbation.

To further understand the effects of perturbation, we train three dataset-dedicated, binary classifiers on the training subset of each of our three datasets: landmark, wavelet, and StyleGAN2 datasets. The result is three separate FaceNet classifiers trained to differentiate between morph and genuine images for specific datasets.

We refer to these classifiers as the dedicated FaceNet classifiers. Using a testing subset of each dataset, we test the effects of perturbation on the dedicated FaceNet classifiers' respective datasets. For example, the landmark and landmark perturbed datasets are tested

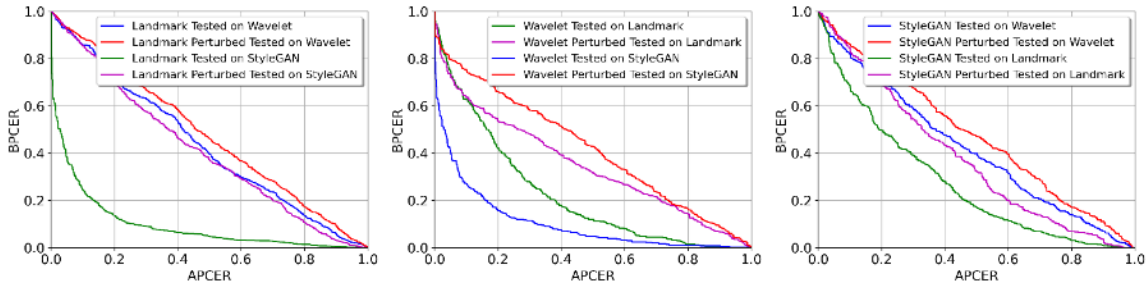

Figure 3.11: Cross dataset testing on the dedicated FaceNet classifiers.

Table 3.7: Cross-dataset APCER and BPCER results from dedicated FaceNet classifiers.

Dataset	**Dedicated FaceNet Classifier**	**AUC**	**APCER@BPCER**			**BPCER@APCER**			**EER**
			1%	**5%**	**10%**	**1%**	**5%**	**10%**	
Landmark	StyleGAN2	90.39	89.53	51.86	28.48	58.12	35.56	23.71	18.90
Landmark Perturbed	StyleGAN2	58.71	97.09	88.13	81.62	97.66	90.96	84.59	43.22
Landmark	Wavelet	50.63	99.30	91.04	85.46	97.33	92.17	85.80	44.67
Landmark Perturbed	Wavelet	52.07	99.65	93.95	89.88	98.47	92.98	88.46	47.66
Wavelet	StyleGAN2	89.65	90.21	51.57	34.54	58.08	37.25	26.22	17.64
Wavelet Perturbed	StyleGAN2	58.14	100.00	94.60	89.21	86.21	79.36	74.11	45.71
Wavelet	Landmark	77.75	85.38	66.90	56.76	89.95	72.30	62.50	28.67
Wavelet Perturbed	Landmark	65.19	99.26	93.62	84.55	87.45	71.76	63.59	39.40
StyleGAN2	Wavelet	58.12	98.60	92.44	86.62	96.44	88.00	80.00	44.20
StyleGAN2 Perturbed	Wavelet	52.28	98.95	95.81	90.96	96.88	91.55	85.11	48.70
StyleGAN2	Landmark	72.70	92.44	76.04	64.18	96.00	78.22	66.00	34.66
StyleGAN2 Perturbed	Landmark	62.23	95.58	88.37	76.74	98.44	90.00	82.22	41.55

on the landmark dedicated FaceNet classifier as presented in Table 3.6 and Figure 3.10. The landmark wavelet dataset has a drop in AUC from 78.63% to 59.10% and an increase in EER from 28.67% to 43.52%. Additionally, the APCER @ BPCER 1% spikes from 90.00% to 98.37%, Across the datasets we see a spike in APCER @ BPCER, meaning a significant jump in morph misclassification. Clearly, the perturbation is causing the classifier to erroneously classify the morphed images, leading to a higher rate of morphs being labeled as genuine and, therefore, an increase in ACPER@BPCER for all thresholds as well as an increase in EER. The results show that the perturbation applied to the morph images has a significant effect on morph detectors optimized for a particular morphing

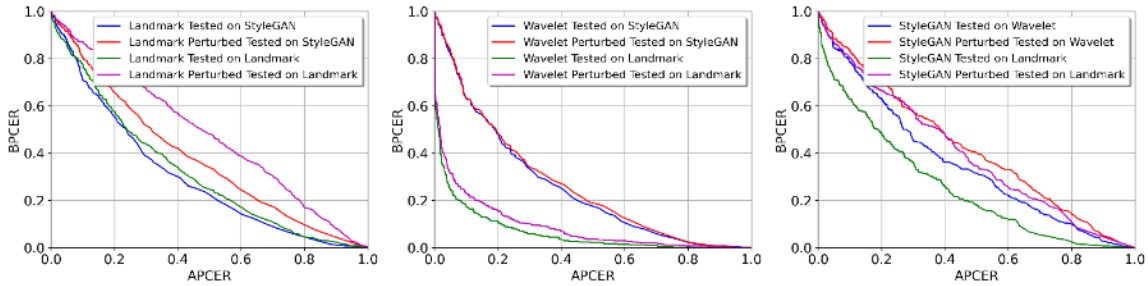

Figure 3.12: Cross dataset testing on the dedicated morph detectors used as verifiers.

Table 3.8: Cross dataset APCER and BPCER results from dedicated morph detectors used as verifiers.

Dataset	**Dedicated Verifier**	**AUC (%)**	**APCER@BPCER**			**BPCER@APCER**			**EER (%)**
			1%	**5%**	**10%**	**1%**	**5%**	**10%**	
Landmark	StyleGAN2	70.16	88.48	74.50	67.79	95.32	85.56	70.48	34.50
Landmark Perturbed	StyleGAN2	61.96	91.40	89.76	84.32	96.61	90.64	79.59	40.56
Landmark	Wavelet	68.32	97.09	80.00	71.74	93.22	84.19	75.72	36.61
Landmark Perturbed	Wavelet	51.97	99.41	93.25	88.25	97.01	92.33	86.20	48.62
Wavelet	StyleGAN2	74.36	87.19	71.13	64.15	94.11	80.63	62.63	32.45
Wavelet Perturbed	StyleGAN2	73.36	87.19	76.82	67.98	94.60	80.88	62.99	34.50
Wavelet	Landmark	93.60	70.28	36.11	21.93	46.81	24.01	17.15	13.72
Wavelet Perturbed	Landmark	91.05	84.42	45.16	33.21	51.96	31.37	22.79	17.15
StyleGAN2	Wavelet	64.54	95.81	86.81	80.58	94.44	84.22	78.22	37.77
StyleGAN2 Perturbed	Wavelet	57.67	98.83	91.04	85.93	97.55	88.44	82.22	44.44
StyleGAN2	Landmark	74.58	85.00	73.25	63.13	84.88	71.11	62.00	32.00
StyleGAN2 Perturbed	Landmark	60.38	89.50	82.70	74.80	95.11	86.00	79.33	42.66

technique. Further, in Table 3.7 and Figure 3.11, we test the cross-dataset performance on the dedicated FaceNet classifiers. For every test, we again see a decrease in AUC and an increase in EER. By testing across the FaceNet classifiers, we show that the added adversarial perturbation is transferable to unseen classifiers. Table 3.7 and Figure 3.11 shows that the landmark dataset tested across detectors is most greatly affected by perturbation.

To further explore the transferability of the designed adversarial perturbations for the morphed images, we use our dedicated detectors as verifiers in a differential morph detection setting. In Table 3.8 and Figure 3.12, the differential morph detection performance of each dedicated detector is tested against images perturbed on other dedicated detectors. Across all datasets, AUC drops and the perturbed StyleGAN2 dataset preserves the best

transferability. Most notably, the AUC of the StyleGAN2 dataset tested on the landmark detector sees a drop in AUC of 14.20%. For all tests, we see an increase in APCER@BPCER 5% of at least 5%, showing a strong increase in attack rates. Again, we show the advantage of adversarially perturbing morph data to improve the quality of their attack. We also observe that, considering Tables 3.7 and 3.8, the adversarial perturbation transfers better for the single morph classification compared to differential morph detection.

3.2.6 Summary

Morphing twins is a significant challenge for FRS that leads to erroneous verification, with our twin datasets scoring over 10% AUC lower than datasets found in literature. We were able to show that the twin morphs represent an extremely difficult scenario for FaceNet, leading to abnormally high error rates. With FaceNet EER values above 30% for all three twin datasets, the need for more work on extreme cases such as twin morphs is emphasized. To further improve the attack quality of our morphed images, we explored the effect of adding adversarial perturbation to our morph datasets. We showed that the perturbation is transferable across several unseen classifiers. The perturbation gave the already difficult twin morph dataset even greater capabilities. The generated twin morphed images are one of the ultimate challenges for an FRS and as such, these images can be used to further test the accuracy of automated morph detectors.

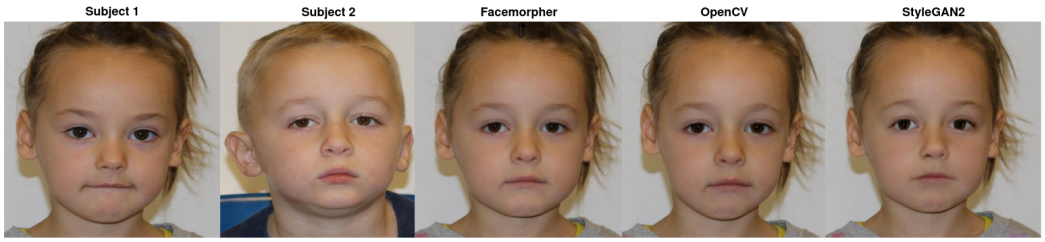

Figure 3.13: The bona fide subjects and morphed samples from Clarkson dataset.

3.3 Kid Morph Generation

Facial recognition systems perform lower on children than adults. Michalski *et al.* show that commercial-off-the-shelf (COTS) algorithms at an FMR of 0.1% in a verification setting for juveniles result in a false match rate up to six times higher than adults [56]. One of the major barriers to the improvement of juvenile face recognition is the lack of publicly available datasets dedicated to children [56]. Most FRS common in literature are trained on large publically-available datasets such as Visual Geometry Group Face2 (VGGFace2) [57]. While these datasets contain children's faces, the proportion of juvenile subjects is statistically insignificant to create reliable FRS for children. Srinivas *et al.* [58] study multiple COTS and government-off-the-shelf (GOTS) algorithms to understand the bias FRS have against children. They were able to deduce that in both identification and verification scenarios, children do not perform as well as adult baselines. Similarly, the Face Recognition Vendor Test (FRVT) [56, 59] has consistently shown lower performance on child subjects than on adult faces.

Additionally, children are more difficult to verify in person than adult subjects, creating a challenge for in-person verification which would otherwise come naturally [60]. We propound this crucial scenario with serious implications for national security and child trafficking: If a bad actor attempts to cross an international border with a child, the bad actor can create a morphed image of the child with a look-alike and pass the child through

Table 3.9: Morph detection performance on our six morphed datasets.

	Morph Dataset	AUC	APCER@BPCER			BPCER@APCER			EER
			1%	5%	10%	1%	5%	10%	
Differential	Clarkson StyleGAN2	89.75%	46.55%	36.72%	28.85%	72.86%	55.85%	32.44%	16.73%
	Clarkson OpenCV	83.58%	58.57%	51.13%	44.33%	76.32%	67.45%	48.81%	24.74%
	Clarkson Facemorpher	83.86%	54.85%	47.76%	40.29%	75.68%	71.13%	52.97%	24.70%
	UNCW StyleGAN2	97.32%	27.22%	13.70%	5.05%	42.40%	15.00%	8.51%	9.44%
	UNCW OpenCV	92.23%	48.15%	28.97%	18.03%	77.66%	44.97%	30.17%	14.64%
	UNCW Facemorpher	89.45%	53.90%	40.75%	30.78%	80.52%	51.45%	37.20%	18.45%
Single	Clarkson StyleGAN2	79.68%	86.57%	69.57%	52.47%	97.91%	71.00%	55.39%	27.06%
	Clarkson OpenCV	69.89%	93.15%	75.29%	65.60%	99.52%	91.77%	78.48%	36.12%
	Clarkson Facemorpher	70.47%	94.52%	74.33%	64.78%	97.76%	92.00%	80.56%	33.54%
	UNCW StyleGAN2	92.66%	72.35%	29.39%	21.44%	71.15%	40.94%	26.17%	14.55%
	UNCW OpenCV	81.38%	95.34%	76.02%	58.12%	74.07%	59.39%	44.24%	24.77%
	UNCW Facemorpher	81.11%	95.59%	73.32%	58.88%	73.25%	60.01%	45.17%	27.29%

border security under the doppelganger's alias. In 2019, in the United States alone, there were over 6,000 reported cases of adults crossing a border with a minor fraudulently labeled as their own [61]. Our work is vital to detecting vulnerable children in these scenarios.

To the best of our knowledge, this is the first attempt to morph juvenile subjects to create morphed faces. We generate and evaluate 52,686 high-quality morph images utilizing two landmark-based and one generative adversarial morph method for children of the wide age range of 4 to 17 years old. Examples of our generated morphs from each of our morphing techniques can be found in Fig. 3.3. These images present a difficult scenario for face verification systems and can be utilized to improve FRS models, as well as shed light on the current dangers of morphing children's faces. Many deep learning models show a strong bias against children [59], by morphing children we take advantage of this bias in order to further fool facial recognition systems.

Two datasets are utilized to create our morphed images, the Clarkson University children dataset [62] and UNCW MORPH age-progression dataset [63]. For clarity, we refer to the two datasets as Clarkson and UNCW for the remainder of the section. The two datasets

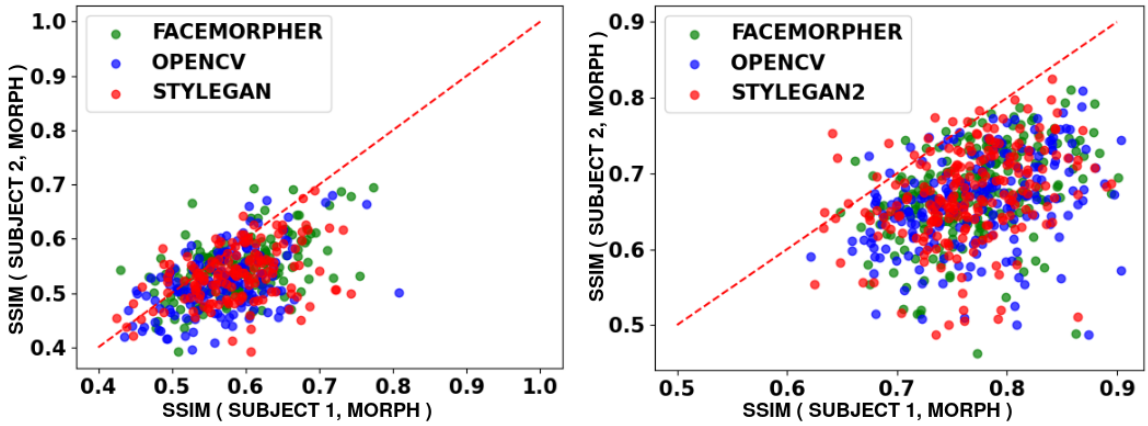

Figure 3.14: SSIM between scores between bona fide and morphed images for the UNCW (left) and Clarkson (right) datasets. The red line indicates the ideal SSIM scores, where both subjects equally contribute to their respective morphs.

are utilized in order to generate a range of generated ages, with the Clarkson dataset containing images of children ages 4-11 years old, and a subset of the UNCW dataset containing subjects ranging from 16-17 years old. For both datasets, the subjects are in front of a neutral background and looking directly into the camera. A four year old has vastly different facial features than a 17 year old. When morphing, it is vital to morph subjects who look-alike in order to reduce morphing artifacts. Therefore, we preserve the integrity of the demographics of each dataset by generating two separate morphed datasets from the respective bona fide datasets.

UNCW dataset: From [63], we extract individuals of age 16-17 years old. The dataset has a strong gender bias, and our subset includes 499 male and 58 female subjects. The images are of size 470×400. Compared to Clarkson dataset, the subjects in this dataset have highly distinguishable features, similar to adults. We use the L_2 distance between the FaceNet's embeddings of length 512 in order to generate a similarity scores [46]. Morphs are generated within gender groups, and similarity scores are calculated within genders. As presented in Fig. 3.15, distance threshold is set at the top 5% of the female pairs in the distribution and pairs below this threshold are considered look-alikes. This threshold is

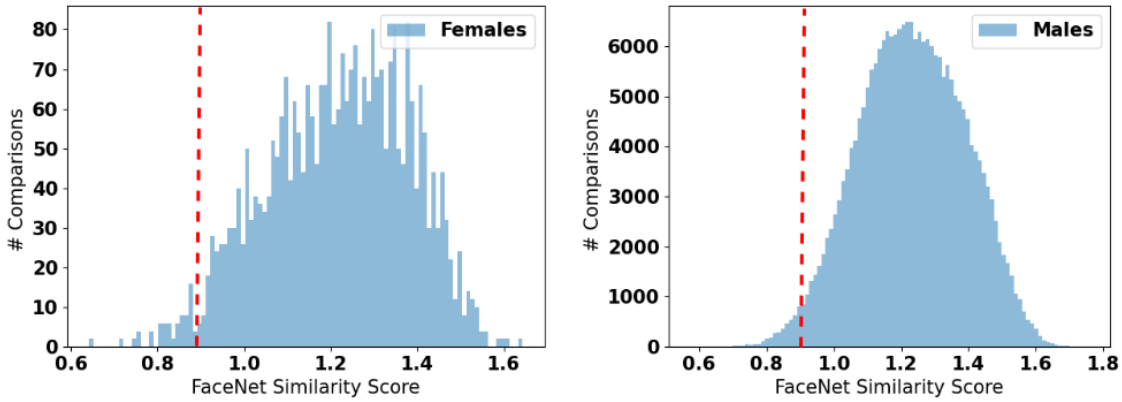

Figure 3.15: All-to-all distribution for comparisons of subjects from the UNCW dataset. Pairs below the distance threshold are considered look-alikes.

also applied to the male distribution. 465 subjects are accounted in the final pairings, and per morphing method, 7,564 morphs are generated.

Clarkson dataset [62] is made up of children ages 4-11 years old. The original images are of sizes 5472×3648 and of good visual quality. We used a subset of the data containing 165 subjects. The children are so young their faces lack highly distinguishable features, thus, creating high inter-class similarity between the subjects. Therefore, using FaceNet we find the top 10,000 look-alike pairs and use them for morphing. We cropp the images to 512×512 and morph using the Facemorpher landmark-based, OpenCV landmark-based, and StyleGAN2 techniques. The resulting images are 512×512 and have no visual morphing artifacts.

3.3.1 Vulnerability Analysis

Morphed images contain structural similarities with their bona fide subjects. Structural Similarity Index Measure (SSIM) [42] is calculated based on perceived similarity between reference images rather than a pixel-to-pixel comparison. As presented in Fig. 3.3, we compare the SSIM score between the bona fide identities and their respective morphs. A

Table 3.10: MMPMR (%) for our six juvenile datasets.

Method		Facemorpher	OpenCV	StyleGAN2
Clarkson	FaceNet	91.31	87.98	73.82
	ArcFace	90.02	83.80	62.45
UNCW	FaceNet	99.32	97.87	90.40
	ArcFace	97.25	93.13	81.49

higher SSIM score represents greater structural similarity. The datasets show a linear correlation between the structural similarities of bona fide identities and the morphed image. While the Clarkson dataset's SSIM scores trend higher than UNCW, it has a higher variance. Meaning, the Clarkson dataset maintained similarity better than UNCW, but shows greater bias toward one contributing subject over another. This is due to the greater variable face shapes in the young children in the Clarkson dataset. Therefore, when the convex hull is placed onto a contributing subject's face the SSIM is biased toward the subject used as the background of the morphed face, *i.e.*, the image with stronger structural similarities.

Again, we use both MMPMR and APCER/BPCER as out metrics of comparison. As presented in Table 3.10, we use FaceNet [46] and ArcFace [52] as our verifiers with τ as the operational verification threshold at False Match Rate (FMR) of 0.1% [53]. For our six juvenile morph datasets, FaceNet is more vulnerable compared to ArcFace. In addition, the landmark morphing datasets provide higher vulnerability compared to StyleGAN2 datasets. This observation is consistent with previous studies on landmark- and StyleGAN-based morph generation [21].

3.3.2 Differential morph detector

We use FaceNet [46] as a verifier for our morphed images as shown in the Table **??**. We consider a positive pair a genuine image of a subject paired with a secondary bona fide instance of the subject, while a negative pair is a genuine image paired with a subject's re-

spective morph. Verification results with a lower Area Under the Curve (AUC) and higher APCER values indicate that the morphs are successfully fooling the verifier. The morphed childrens' faces are able to fool FaceNet, with Equal Error Rate (EER) values over 9%. Across the three methods of morphing, StyleGAN2 consistently has a higher AUC then the landmark-based morphs. For example, while the Clarkson StyleGAN2 dataset has an AUC of 89.75%, the OpenCV and Facemorpher versions of the dataset have AUC 83.58% and 83.86%, respectively. This trend implies that FaceNet is able to differentiate between the morph and bona fide StyleGAN2 images at a higher rate than the landmark morph datasets. These results reinforce the known issue that StyleGAN2-generated morphs struggle to retain identity information at the same rate as the landmark-based morphs [12].

The verification results for the landmark morph dataset are significantly lower than FaceNet's expected morph detection performance. In [13], adult datasets are verified over 99% AUC using FaceNet. The morphed child datasets results in a significant AUC drop of approximately 16% when compared to adults. Additionally, there is a significant difference in performance of the verifier when comparing the older children in the UNCW and the young children found in the Clarkson dataset especially using the OpenCV method where the Clarkson OpenCV dataset has an AUC of 83.58% and the UNCW OpenCV dataset with an AUC of 92.23%.

3.3.3 Single morph Detection

Using FaceNet [46], we train a binary classifier with a two-node output to detect morphs. The morph detector is trained on approximately 12,000 Facemorpher, OpenCV, and StyleGAN images of adult datasets. The detector learns the common artifacts of images using these morphing techniques. Table **??** shows the performance of the classifier on our six juvenile datasets. A lower AUC and higher APCER value indicate stronger morphed images

because the morph images are fooling the classifier. Similar to the differential scenarios, StyleGAN2 is shown to have a higher AUC in classification than the other datasets, specifically having an AUC of 79.68% for the Clarkson StyleGAN2 dataset and 92.66% AUC for the UNCW StyleGAN2 dataset, while their respective landmark morphs trend approximately 10% lower. This means that the SyleGAN2 datasets have more artifacts than their respective landmark-based datasets. The Clarkson landmark morphs and the UNCW landmark morphs all have APCER at BPCER=1% values above 93%, meaning that the morphs are effective at fooling the morph detector. In this scenario, we again observe the effects of aging in the performance of the classifier. The Clarkson dataset has a higher EER and lower AUC when across the methodologies. For the OpenCV morphs, Clarkson has an EER of 36.12% while UNCW has an EER of 24.77%. For Facemorpher, the EER for Clarkson is 33.54% and UNCW has an EER of 27.29%. This trend continues with StyleGAN2 having an EER of 27.06% and 14.55% for Clarkson and UNCW, which illustrates a bias toward the older children.

3.3.4 Summary

In this paper, we generated high-quality morphed images from juvenile subjects. The morphed images were shown to retain their identity while being convincing enough to fool both single and differential morph detectors. While all datasets are shown to be effective at fooling morph detectors, the landmark-based morph images were more effective compared to StyleGAN2 morphs, which is consistent with adults datasets generated with the same methodology [12]. Across all morph detectors, morphed children pose a more significant threat than adult morphed datasets because of inherent bias when training deep learning models. This illustrated the necessity of further work to bridge the gap between facial recognition in adults and children as juvenile morphed images remain a threat to national security and child safety.

Chapter 4

Conclusion

4.1 Limitations

This work was limited by the both the number of publicly available datasets and morph generation techniques. There are few datasets used in literature that contain images of passport quality. Most large-scale datasets are comprised of faces in-the-wild which are not ideal for morphing. Further, morphing requires intimate knowledge of subject identities for testing purposes. For example, differential morph detection requires a minimum of two images per subject. Many large-scale datasets do not contain adequate documentation of their subjects, normally this is done for privacy reasons. Fortunately, for the majority of our work we were able to turn to private datasets which includes detailed identity documentation. Further, morph generation is constrained to either Landmark or GAN-based morph generation techniques. Of these methods, only a handful of algorithms are available to the public. In order to generate more convincing morphed images, more repositories must be published in order create greater flexibility in the generation phase.

4.2 Next Steps

The natural next step of this work is to train morph detectors on the datasets. Because our morphs are challenging to detect, they are a good training set for morph detection deep learning models. Further, because our datasets are dedicated to specific groups of individuals, morph detectors can be trained to specialize in specific groups of morphed images. For example, for the first time, there is the prospect of training a morph detector to specialize in detecting morphed images of children.

Outside the scope of morph detection, the generated morphed images can be used to test the effectiveness of face verifiers. An ideal FRS will be able to differentiate a morph from a genuine image of a given subject for every instance. The Twin Morph dataset is perfect for this scenario. FaceNet is an extremely accurate face verifier, and the Twin Dataset was still able to achieve EERs over 20%. By training a face verifier with twins, it can be hypothesized that the verifier would be extremely effective.

4.3 Conclusion

In this work we were able to generate high-quality morphed images that successfully fooled FRS. We took advantage of the lack of under-represented groups of individuals in order to create the most challenging morphed images possible. We showed that morphed images are able to successfully fool facial recognition systems into erroneously verifying an individual as a bona fide subject. Further, we show that when adversarial perturbation is applied to images, it can be transferred across morph detectors. In a real-world scenario, these morph images would likely be accepted as genuine. In closing, our high-quality morphed images are likely to fool both face recognition algorithms and humans alike.

Bibliography

[1] "Biometrics," 2021. [Online]. Available: https://www.dhs.gov/biometrics

[2] L. R. Carlos-Roca, I. H. Torres, and C. F. Tena, "Facial recognition application for border control," in *2018 International Joint Conference on Neural Networks*, 2018, pp. 1–7.

[3] ICAO, "9303- machine readable travel documents part 9: Deployment of biometric identification and electronic storage of data."

[4] B. Chaudhary, P. Aghdaie, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Differential morph face detection using discriminative wavelet sub-bands," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1425–1434.

[5] P. Aghdaie, B. Chaudhary, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi, "Detection of morphed face images using discriminative wavelet sub-bands," *CoRR*, vol. abs/2106.08565, 2021. [Online]. Available: https://arxiv.org/abs/2106.08565

[6] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] I. D. Raji and G. Fried, "About face: A survey of facial recognition evaluation," *arXiv preprint arXiv:2102.00813*, 2021.

[8] C. Boutin, "NIST evaluation shows advance in face recognition software's capabilities," 2018.

[9] A. Quek, "Facemorpher," Jan 2019. [Online]. Available: https://github.com/alyssaq/face$_morpher$

[10] S. Mallick, "Face morph using OpenCV - C++/Python," March 2016. [Online]. Available: https://learnopencv.com/face-morph-using-opencv-cpp-python/

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

[12] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," *CoRR*, vol. abs/2012.05344, 2020. [Online]. Available: https://arxiv.org/abs/2012.05344

[13] K. O'Haire, S. Soleymani, B. Chaudhary, P. Aghdaie, J. Dawson, and N. M. Nasrabadi, "Adversarially perturbed wavelet-based morphed face generation," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 01–05.

[14] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019, pp. 4401–4410.

[15] P. Phillips, H. Wechsler, J. R. Huang, and P. J. Rauss, "The FERET database and evaluation procedure for face-recognition algorithms," *Image Vis. Comput.*, vol. 16, pp. 295–306, 1998.

[16] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 947–954 vol. 1.

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

[18] L. DeBruine, “Face research lab london set,” January 2017.

[19] R. Abdal, Y. Qin, and P. Wonka, “Image2StyleGAN: How to embed images into the stylegan latent space?” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4432–4441.

[20] “Passport photos.” [Online]. Available: https://travel.state.gov/content/travel/en/passports/how-apply/photos.html

[21] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “MIP-GAN—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.

[22] N. Damer, A. M. Saladie, S. Zienert, Y. Wainakh, P. Terhörst, F. Kirchbuchner, and A. Kuijper, “To detect or not to detect: The right faces to morph,” in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[23] 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,” in *International Conference of the Biometrics Special Interest Group*, 2017, pp. 1–7.

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” 2014.

[25] S. Venkatesh, H. Zhang, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “Can GAN generated morphs threaten face recognition systems equally as landmark based morphs?-vulnerability and detection,” in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, 2020, pp. 1–6.

[26] 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.

[27] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[28] Q. Liao, Y. Li, X. Wang, B. Kong, B. Zhu, S. Lyu, Y. Yin, Q. Song, and X. Wu, "Imperceptible adversarial examples for fake image detection," *CoRR*, vol. abs/2106.01615, 2021. [Online]. Available: https://arxiv.org/abs/2106.01615

[29] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3348–3357.

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

[31] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[32] M. Sharif, S. Bhagavatula, L. Bauer, and M. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[33] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188–5196.

[34] U. Scherhag, C. Rathgeb, and C. Busch, "Morph deterction from single face image: A multi-algorithm fusion approach," ser. ICBEA '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 6–12. [Online]. Available: https://doi.org/10.1145/3230820.3230822

[35] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[36] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *International Conference on Biometrics*. Springer, 2007, pp. 828–837.

[37] G. Lowe, "Sift-the scale invariant feature transform," *International Journal of Computer Vision*, vol. 2, no. 91-110, p. 2, 2004.

[38] 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.

[39] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[40] L. DeBruine and B. Jones, "Face research lab london set," January 2018.

[41] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015. [Online]. Available: http://arxiv.org/abs/1503.03832

[42] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[43] J. R. Paone, P. J. Flynn, P. J. Philips, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, M. T. Pruitt, and J. M. Grant, "Double trouble: Differentiating identical twins by face recognition," *IEEE Transactions on Information forensics and Security*, vol. 9, no. 2, pp. 285–295, 2014.

[44] 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.

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

[46] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[47] J. McCauley, S. Soleymani, B. Williams, J. Dando, N. Nasrabadi, and J. Dawson, "Identical twins as a facial similarity benchmark for human facial recognition," in *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2021, pp. 1–5.

[48] S. Biswas, K. Bowyer, and P. J. Flynn, "A study of face recognition of identical twins by humans," *2011 IEEE International Workshop on Information Forensics and Security*, pp. 1–6, 2011.

[49] I. O. for Standardization, "ISO/IEC DIS 30107-3:2016: Information technology biometric presentation attack detection," p. 3:Testing and Reporting, 2017.

[50] T. Neubert, A. Makrushin, M. Hildebrandt, C. Krätzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biom.*, vol. 7, pp. 325–332, 2018.

[51] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper, "Pw-mad: Pixel-wise supervision for generalized face morphing attack detection," in *International Symposium on Visual Computing*. Springer, 2021, pp. 291–304.

[52] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[53] Frontex, "Best practice technical guidelines for automated border control (ABC) systems," 2015.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *AAAI Conference on Artificial Intelligence*, 2017.

[55] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[56] D. Michalski, S. Y. Yiu, and C. Malec, "The impact of age and threshold variation on facial recognition algorithm performance using images of children," in *International Conference on Biometrics*, 2018, pp. 217–224.

[57] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *IEEE international conference on automatic face & gesture recognition*, 2018, pp. 67–74.

[58] N. Srinivas, K. Ricanek, D. Michalski, D. S. Bolme, and M. King, "Face recognition algorithm bias: Performance differences on images of children and adults," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 2269–2277.

[59] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing face recognition vendor test (FRVT) part 1: Verification," *National Institute of Standards and Technology*, 2018.

[60] D. Kuefner, V. Macchi Cassia, M. Picozzi, and E. Bricolo, "Do all kids look alike? evidence for an other-age effect in adults." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 34, no. 4, p. 811, 2008.

[61] J. Davis, "Border crisis: CBP fights child exploitation," 2020. [Online]. Available: https://www.cbp.gov/frontline/border-crisis-cbp-fights-child-exploitation

[62] P. Das, L. Holsopple, D. Rissacher, M. Schuckers, and S. Schuckers, "Iris recognition performance in children: A longitudinal study," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, pp. 138–151, 2021.

[63] K. Ricanek and T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," in *IEEE international conference on automatic face and gesture recognition*, 2006, pp. 341–345.