StyleGAN and StyleGAN2 for Face Morphing Applications

Jason Kuo

Abstract

Face morphing presents a possible attack on face recognition systems by allowing multiple faces to successfully authenticate for a single reference face. We offer a modularized pipeline in which these recognition algorithms can be evaluated against many different morphing methods. We also explore the performance of StyleGAN and StyleGAN2 for generating morphs using this pipeline as well as how they compare with more traditional landmark-based morphs. Our evaluation indicates that, for a popular open source face recognition algorithm, these GAN-based methods are less deceptive than the landmark-based ones and therefore do not yet pose as a threat to face recognition systems. We also discuss future directions that could allow the state of GAN-based morphing to improve.

1 Introduction

Face recognition is an important feature in many systems today, including unlocking personal devices and verifying identity. One particular application of note is an eGate system which is used to automatically permit authenticated users into a restricted area such as for boarding a plane [1]. Face recognition, along with other biometric verification systems, function by comparing the known data about an individual (such as their face on a passport image) to new data at the gate (for example, by taking a live picture with a webcam).

Face morphing poses a potential security threat to systems like these because if a morph is successfully submitted as a passport photo, it could authenticate multiple users with the same image [1]. There are several methods for generating these morphs. Traditionally, landmark-based morphing methods have been used to detect important facial landmarks and interpolate between the corresponding landmarks for each face contributing to the morph (usually only 2 faces). The highest quality versions of these methods require manual placement of landmarks, which is extraordinarily slow and labor intensive. This process can be automated, but many of the current fully automatic landmark-based algorithms produce easily seen morphing artifacts.

A more modern approach is to use Generative Adversarial Networks (GANs) to produce morphs. In general, GANs attempt to generate new data from the underlying distribution of input data [2]. They do this by constructing two neural networks, a generator and a discriminator, and setting them against each other in a zero-sum game. The generator's goal is to learn to generate data samples similar to the input data for the purpose of fooling the discriminator, whereas the discriminator's goal is to learn to distinguish which samples come from the actual input and which come from the generator. After training, the generator with its learned model should be able to produce data similar to the input, but not by simply copying the input. For the specific example of face images, GANs usually train two convolutional neural networks. Most GANs produce the output image starting with some latent vector representation, meaning that we can think of a GAN as mapping an input vector to a face image and training the GAN is just learning that mapping. For a given input face, if we can find the representation such that the out-

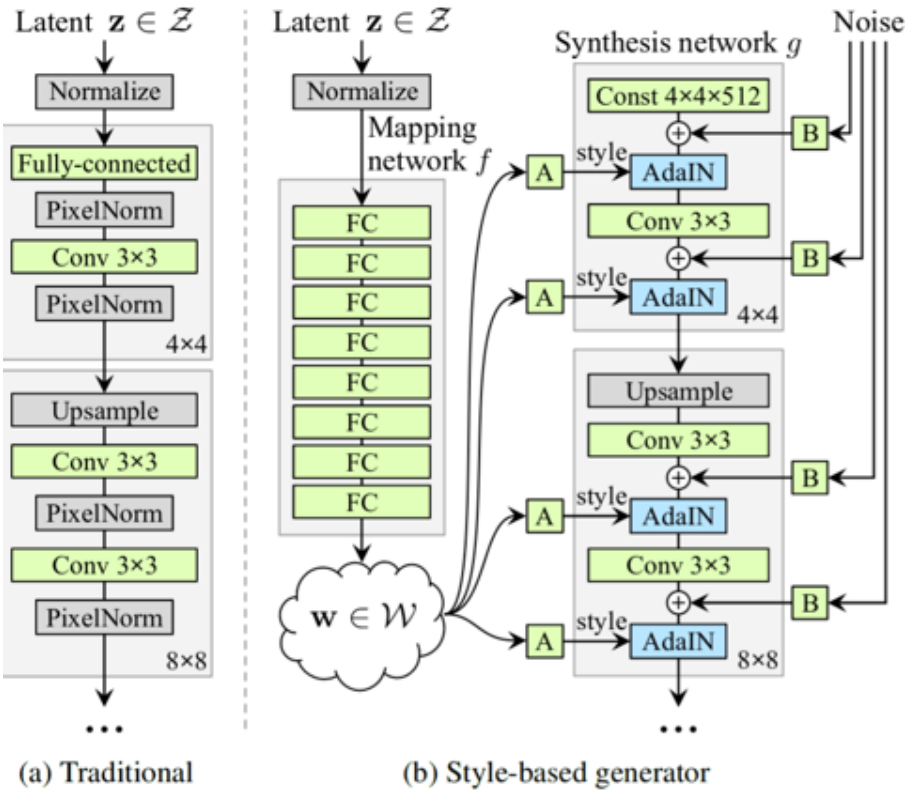


Fig. 1: The style-based architecture of StyleGAN maps a latent vector $\mathbf{z}$ to an intermediate latent vector $\mathbf{w}$ whereas a traditional generator only feeds the latent vector into the input layer. StyleGAN then feeds the intermediate $\mathbf{w}$ into different layers for each resolution to produce a high resolution result.

put after GAN processing is as close to the input face as possible, then we can simply interpolate between two such representations to generate morphs between the two faces that those vectors represent. Interpolation is meaningful [3] in the context of GANs because they attempt to learn a smooth mapping from the latent space of representations to the output images.

In this paper, we will try to evaluate some GAN-based morphing methods and compare them with a landmark-based method. We will primarily focus on modifying and using StyleGAN and StyleGAN2 to generate these GAN-based morphs.

2 StyleGAN

StyleGAN [4] is a GAN architecture (Figure 1) from NVIDIA which is able to generate high quality and high resolution images. It is inspired by style transfer literature which emphasizes the importance of meaningful interpolation properties. Thus, it is very useful for morphing applications. By default, as well as in our pipeline, StyleGAN is trained on the FFHQ dataset consisting of 70,000 high-quality face images at 1024x1024 resolution collected from Flickr. However, one problem with the original StyleGAN is that it does not provide a means for finding the latent vector that corresponds to a given input image. That is, the inverse mapping that converts from a face image to a latent vector is not learned with NVIDIA's implementation.

A popular implementation that does provide this functionality is the stylegan-encoder repository [5]. It uses a pre-trained VGG16 network to calculate the feature vectors for the given input image and an initial StyleGAN generated image. A loss function is defined as the difference between these two vectors. Then, it minimizes this loss by optimizing the generated image through the initial latent vec-

(a) Subject A (b) Landmark (c) StyleGAN (d) StyleGAN2 (e) Subject B

Fig. 2: Visual comparison of various morphing methods. Note that both of the GAN-based morphs require a face-aligned image which can be a differentiating feature compared to many landmark-based methods. It is also clear in this comparison that StyleGAN produces distortions and other visual artifacts (especially near the border) whereas StyleGAN2 does not.

tor to obtain the best encoded latent representation for the input image.

3 StyleGAN2

More recently, NVIDIA released an improved version of StyleGAN named StyleGAN2 [6]. It builds on the original by focusing on removing normalization artifacts which take the form of blurry blobs in the image. It also improves the image quality and provides built in projection functionality for mapping from an image to its latent representation. Their projector mainly differs from the third-party stylegan-encoder by focusing on finding latent representations for images that the StyleGAN2 generator would have produced on its own. In constrast, the details of the stylegan-encoder's methodology ends up extending the latent space, which allows encoding arbitrary images that would not otherwise have a latent representation. We will see that this crucial difference impacts the morph results and our evaluation of them.

4 Experimental Methods

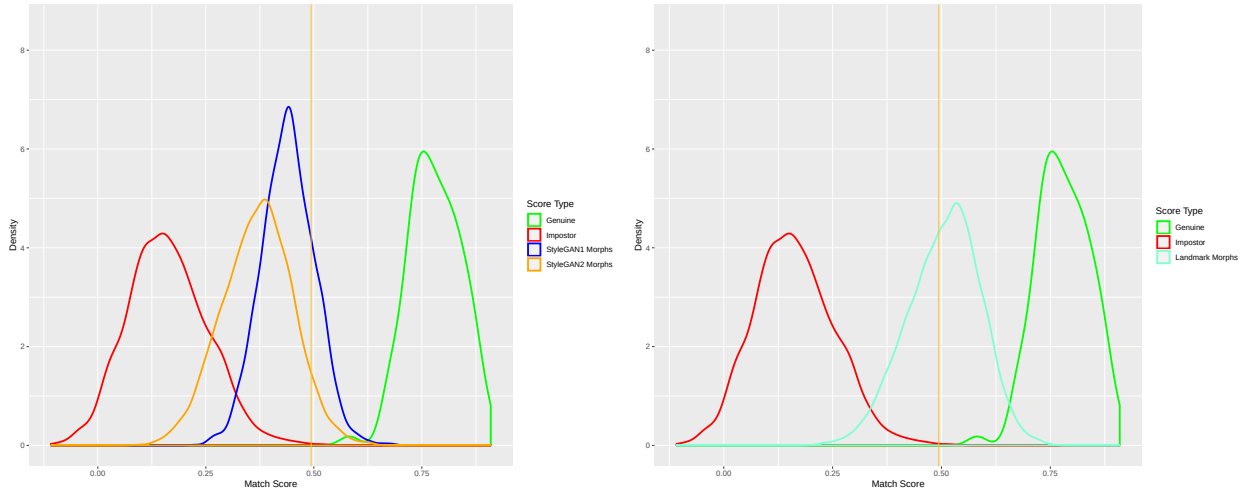
Our goal is to evaluate to what extent a face recognition algorithm is fooled by GAN generated morphs. To this end, we set up a pipeline for generating morphs, feeding them into a face recognition algorithm, and then analyzing the recognition scores produced.

4.1 Setup

We implemented this pipeline on a CentOS 7.6 machine with 2 Intel Xeon CPUs and 2 Tesla K40m NVIDIA GPUs. We used the Face Research Lab London Dataset [7] for all of our testing. This dataset was separated into male and female faces so that morphs would only be generated within gender. Then, using both the stylegan-encoder and the StyleGAN 2 projector, 50-50 morphs (equal contribution morphs generated using 0.5 as the interpolation coefficients) were generated between each pair of faces within each gender group.

These morphs were evaluated using the `face_recognition` Python module which is a wrapper around Dlib's face recognition features. This module produces a face distance d which is then inverted with $1 - d$ to produce a similarity score for our evaluation. We compared each morph with the two images from which it is generated to produce the morph distribution. We also generate a genuine distribution by comparing between the neutral and smiling versions of each face and an impostor distribution by comparing between neutral images of different people.

As another baseline, we also generated landmark-based morphs by automatically detecting important landmarks of the input faces, averaging them, and splicing them onto one of the initial faces as described by [8]. During evaluation, these landmark morphs were also compared with each of the two images that went into it.



(a) Plot of morph score distributions

(b) Plot of landmark morph scores

Fig. 3: The vertical golden line indicates the score threshold for 0.1% false match rate, meaning that only 0.1% of impostors would trick this algorithm if it used this score threshold for classification.

4.2 Results

Figure 3a shows the score distributions for the GAN-based morphs generated using the Dlib-based face recognition algorithm. The vertical line represents an important threshold because 0.1% false match rate (FMR) is a common metric used to calibrate face recognition systems [9]. Since FMR is only calculated based on our invariant impostor distribution, this threshold will be constant at a score of 0.494 for the rest of our analysis. The median score for the stylegan-encoder morphs is 0.440 whereas the median score for the StyleGAN2 projector is 0.372. Similarly, figure 3b shows the score distribution for our landmark-based algorithm. The median landmark morph score is higher than both of the GAN-based ones at 0.512.

The proportion of the stylegan-encoder morphs that are above the threshold was 0.192. For the StyleGAN2 projector that becomes 0.059. For the landmark morphs, it is 0.580.

Learning the latent vectors through the stylegan-encoder took about 6 minutes per face whereas doing the same with the StyleGAN2 projector took about 20 minutes per face. After this process, generating morphs took only a couple seconds per morph.

5 Discussion

From our evaluation, it seems that the stylegan-encoder and the StyleGAN2 projector are not feasible for generating highly deceptive morphs yet. In order for a method to successfully trick a face recognition algorithm, the match score of the morph must be greater than the threshold the algorithm uses. For dlib-based algorithm and the common threshold of 0.1% FMR, the algorithm would only be fooled 19.2% of the time by the StyleGAN based method. For the StyleGAN2 based method, this is even worse at about 6%. This contrasts with the landmark-based morphs (at 58%) which are already unreliable for consistently fooling the face recognizer.

The cause of the low scores likely comes from the latent learning process. Figure 4 shows the match score distributions for the images generated directly from the latent vectors before any morphing takes place. Since no morphing occurs at this stage, each latent image should represent the same face and ideally have just as high of a match score as the genuine comparisons. However, the scores tend to be lower for the stylegan-encoder latents and much lower for the StyleGAN2 projector latents (only 76.5% of the StyleGAN2 latents

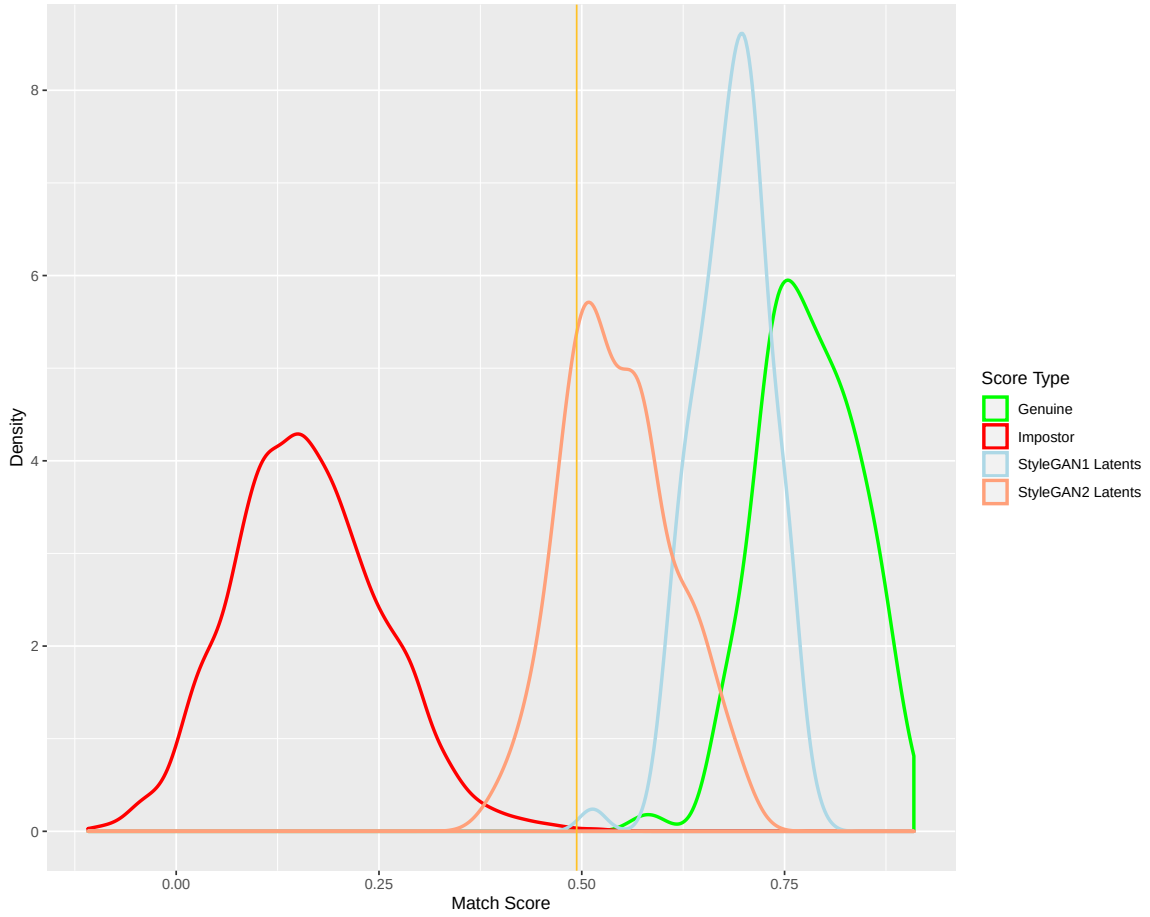


Fig. 4: Plot of the latent score distributions for each GAN-based method. The genuine and impostor distributions are the same as before. This plot shows that the images generated from latent representations without morphing already suffer from flaws.

surpass the threshold). Even though the stylegan-encoder latents all surpass the threshold like the genuine distribution, having lower match scores (median of 0.690 vs. 0.776 for genuines) means that there are slight inaccuracies which will propagate into the morphs.

It is quite interesting that the StyleGAN2 morphs and latents perform much worse than their stylegan-encoder counterparts even though StyleGAN2 is supposed to be of higher quality. This is likely due to the design of the StyleGAN2 projector. Recall that NVIDIA wanted the projector to focus on finding representations that the generator itself would have produced, whereas the stylegan-encoder extends the latent space to allow for encoding of arbitrary images. As a result, the images generated by the stylegan-encoder latents are actually much closer to the original image compared to the StyleGAN2 projector [6].

However, the latent representation is not the only problem. Conceptually, the morphing alone would already decrease match scores since we are trying to find a morph image that is somewhere in between two different faces. It is unclear whether it is possible to find a single face that would look exactly like two different distinct faces or what the upper bound on the similarity can be. Conversely, we can generate higher scoring morphs if the input faces for the morph are already similar. The more similar the inputs are, the higher our morph scores will be since there's less of a transition between the two.

6 Related Work

Abdal *et al.* [3] was one of the first to explore latent embedding interpolation as a method for morphing images using GANs. Their work primarily focused on gaining valuable insight into the structure of the StyleGAN latent space, leading to an efficient algorithm for generating latent embeddings (Image2StyleGAN) and an evaluation of how meaningful the embeddings are. Similarly to the stylegan-encoder, they generate latent space embeddings by optimizing the latent vector to minimize the loss between the given image and the image generated from this latent vector.

Building on their previous work, Abdal *et al.* recently developed Image2StyleGAN++ [10] which improves quality by preserving high frequency features and adds capabilities for local image editing of the embedded images. The embedding generated images which they show in their paper impressively capture minute details from the original given image. If such an embedding process can achieve high match scores for the latent images, then their algorithm may lead to great advancements for GAN generated morphs.

[11] performed a similar evaluation of StyleGAN's morphing abilities as we did. However, they built their own system for embedding and performed their evaluation using commercial face recognition systems and morphing attack detection schemes. Their approach for latent space embedding is similar to the stylegan-encoder since they also optimize the latent vector in order to minimize the loss between the input image and the generated image. Regardless, they reach the same conclusion that GAN generated morphs present less of a threat compared to landmark-based morphs.

[12] extends on [11] by formulating a new loss function which takes perceptual quality and identity prior into account. They referred to this as MIPGAN (Morphing through Identity Prior driven GAN). Using this new model, they consistently perform on par with landmark-based morphs across multiple face recognition systems which theatens any systems without morph attack detection.

[13] presented another evaluation of face morphing attacks using methods similar to [11]. One similarity they had with our paper was that they also used the London dataset. Again, they determined that the morphs generated with StyleGAN2 were not significant threats.

7 Future Work

NVIDIA released StyleGAN2-Ada [14] on October 2020, improving on StyleGAN2 by introducing an adaptive discriminator augmentation mechanism which improves training with small datasets. This yet again improves the quality of the generated images which may provide better results in our evaluation pipeline. Initial testing using our system inidicates that morphs take around 10 minutes for projecting input faces which is twice as fast as the original StyleGAN2. They also released a PyTorch implementation which claims to produce the same results as the original Tensorflow implementation even faster. We hope to perform full evaluations with both of these to see if they produce morphs faster or better.

We would also like to perform our evaluation with multiple face recognition algorithms including some commercial algorithms. This would allow us analyze how the morphs would perform in other circumstances including more realistic environments.

In order to improve the face recognition scores of the latents, another direction we would like to try would be to optimize the stylegan-encoder using the facial recognition scores. This is similar to some of the related work such as [11] and [12]. If the latent scores can be improved with this process, it would likely improve the overall morph distribution as well.

Finally, it would be interesting to look into morphs of already similar looking faces to see how good the morph scores can be. This could let us explore an upper bound on the morphing process. From an adversarial point of view, one would have the best chance of exploiting such a face recognition system by finding a genuine face

that is already similar to the face trying to pass through the system.

8 Conclusion

This paper evaluated the feasibility of StyleGAN and StyleGAN2's use in generating morphed face images for attacking a popular face recognition system. In doing so, we constructed a pipeline for performing the evaluation which can accomodate more components or new components such as morphing methods or face recognition algorithms. We confirmed that the stylegan-encoder built off of StyleGAN and the projector of StyleGAN2 generated morphs that were not able to trick a common open source face recognition algorithm as well as a landmark-based morphing algorithm could. With this pipeline, we hope to evaluate other GAN-based or landmark-based morphing algorithms while incorporating better and more utilized face recognition systems.

References

[1] Matteo Ferrara, Annalisa Franco, and Davide Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space?, 2019.

[4] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[5] Dmitry Nikitko. Stylegan - encoder for official tensorflow implementation, 2019.

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.

[7] Lisa DeBruine and Benedict Jones. Face research lab london set, May 2017.

[8] Andrey Makrushin., Tom Neubert., and Jana Dittmann. Automatic generation and detection of visually faultless facial morphs. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 6: VISAPP, (VISIGRAPP 2017)*, pages 39–50. INSTICC, SciTePress, 2017.

[9] FRONTEX. Best practice technical guidelines for automated border control abc systems, 2015.

[10] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images?, 2020.

[11] Sushma Venkatesh, Haoyu Zhang, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Can gan generated morphs threaten face recognition systems equally as landmark based morphs? – vulnerability and detection, 2020.

[12] Haoyu Zhang, Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Naser Damer, and Christoph Busch. Mipgan – generating strong and high quality morphing attacks using identity prior driven gan, 2021.

[13] Eklavya Sarkar, Pavel Korshunov, Laurent Colbois, and Sébastien Marcel. Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks, 2020.

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.