Approximating Optimal Morphing Attacks using Template Inversion

Laurent Colbois[*1,2], Hatef Otroshi Shahreza[*1,3], Sébastien Marcel[1,2]
[1]Idiap Research Institute, Martigny, Switzerland
[2]Université de Lausanne (UNIL), Lausanne, Switzerland
[3]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
{laurent.colbois,hatef.otroshi,sebastien.marcel}@idiap.ch

Abstract

Recent works have demonstrated the feasibility of inverting face recognition systems, enabling to recover convincing face images using only their embeddings. We leverage such template inversion models to develop a novel type of deep morphing attack based on inverting a theoretical optimal morph embedding, which is obtained as an average of the face embeddings of source images. We experiment with two variants of this approach: the first one exploits a fully self-contained embedding-to-image inversion model, while the second leverages the synthesis network of a pre-trained StyleGAN network for increased morph realism. We generate morphing attacks from several source datasets and study the effectiveness of those attacks against several face recognition networks. We showcase that our method can compete with and regularly beat the previous state of the art for deep-learning based morph generation in terms of effectiveness, both in white-box and black-box attack scenarios, and is additionally much faster to run. We hope this might facilitate the development of large scale deep morph datasets for training detection models.

1. Introduction

Morphing attacks are a particular type of presentation attack which consists in mixing the faces of two contributing subjects to form a so-called *morph*, and submit it as a reference for enrolment in a face recognition system (FRS), for example as a passport photo. In successful attacks, both contributing subjects can then be authenticated by the FRS while using the same passport, which poses an important security issue. While morphing attacks have historically been generated using landmark-based face editing (LMA), several methods have been proposed over the past years that instead exploit deep learning techniques, in particular Generative Adversarial Networks (GANs). The resulting "deep" morphing attacks showcase concerning effectiveness associated with a high realism, although they are generally not yet as successful as LMA ones. Moreover, the most successful GAN-based methods rely on a lengthy latent vector optimization process which renders difficult the generation of a large set of attacks, for example, to create a dataset to train deep morphing attack detection systems.

In parallel of this, recent works in the field of biometric template inversion have demonstrated the feasibility of inverting FRS, i.e., reconstructing a face image starting solely from an extracted face embedding [27, 26]. This has massive implications, as it enables one to perform any kind of

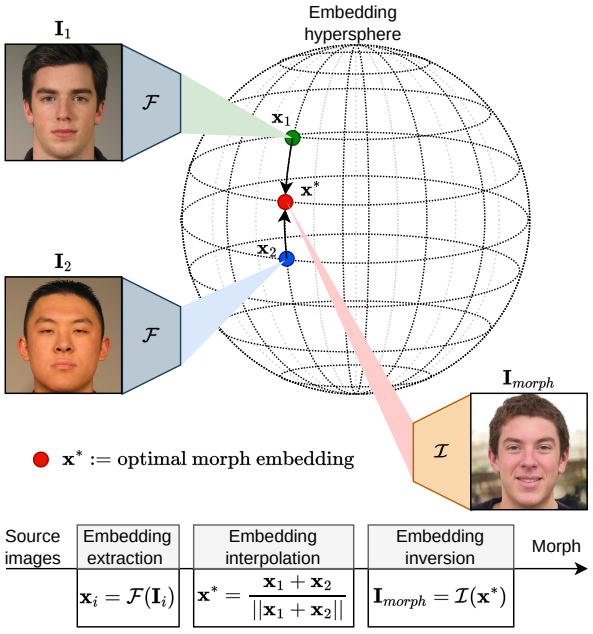


Figure 1. Illustration of the morphing process: face embeddings of the source images are extracted using the FRS $\mathcal{F}$, the corresponding optimal morph embedding is computed by interpolation in the embedding space, then fed back to the template inversion model $\mathcal{I}$ to get the morph.

[*]Equal contribution

arithmetic operations in the embedding space before going back to the image space. In particular, this method can be exploited for morphing attack generation by first computing interpolations between face embeddings of the source identities, then inverting the interpolated embedding back into the image space.

The aim of this work is to evaluate the effectiveness of using a template inversion process for morph generation. Our main contributions are the following:

- we propose a new strategy to generate approximations of the optimal morph images from the corresponding optimal morph embeddings through template inversion techniques,
- we introduce two novel deep morphing attack generation algorithms based on two distinct template inversion methods,
- we evaluate the visual quality of the resulting morphs as well as their attack success rate, both in the settings where the attacked FRS is identical to or different from the inverted FRS.

We observe that the proposed methods are competitive with the previous state of the art in terms of vulnerability, and sometimes even beat it, both in white-box and black-box attack scenarios. Moreover, our morph generation algorithms run orders of magnitude faster than the previous state of the art, making them very practical to generate large deep morphing attack detection datasets.

After contextualizing the state of research on deep morph generation (section 2.1) and template inversion (section 2.2), we detail our novel morph generation methods in section 3.1 and our evaluation process in section 3.2. We then discuss the results both qualitatively (section 4.1) and quantitatively (section 4.2).

2. Related works

2.1. Deep morph generation

Research on morphing attack generation has originally been focused on **landmark-based** methods (LMA). Introduced in [14], those methods proceed by warping the source images to align their facial landmarks, then average pixels between the two warped sources to obtain the morph. As of today, those methods are still typically the most effective at generating morphs able to fool face recognition systems (as evaluated for example in [32]). More recently, new types of morphing generation techniques have arisen, exploiting recent improvements of deep generative models. The idea of using a generative adversarial network (GAN) to generate morphs is first introduced in [9]. Their MorGAN model is obtained by jointly training an encoder from the image space to a latent space, and a generator back from the latent space to the image space. Morph latents are then computed by interpolating between the encoded latents of both source images, then fed to the generator to obtain the morph. [30] and [24] expand on this technique by using instead a pretrained StyleGAN2 network [19]. The image-to-latent encoder is replaced by an optimization process: images are projected in the latent space by finding the latent vector minimizing the perceptual distance between the generated and reference image. Morphs are once again obtained by interpolating the projected latents of the source images and feeding the resulting latent in the generator. With respect to MorGAN, the resulting morphs show significantly improved visual quality, resolution and realism. Several later works take inspiration from those ideas: [8] propose to project LMA morphs in the latent space, before regenerating them with the GAN, in order to get rid of some obvious artifacts. [33] propose similar latent interpolation morphing but replaces the StyleGAN2 backbone by another transformer-based generator architecture. Finally, [32] get rid of the interpolation step, and instead directly explores the latent space in search of a latent whose associated image is an effective morph. This is done by updating the optimization algorithm to take both source images as input, and by including an additional biometric loss that uses a pretrained FRS.

Other types of generative models have also been used: [4] propose a morphing algorithm based on a diffusion process [16]. [20] propose an architecture closer to an autoencoder: a decoder is trained to reconstruct face images from the combination of their face embedding (extracted using some reference face recognition network), as well as from another latent vector encoding all the face image content *not* related to the identity (this encoder is trained jointly with the decoder). Morphs can then be generated by altering the input face embedding fed to the decoder to use instead a worst-case morph embedding between the two source identities. Conceptually, this approach starts from the same goal as our work (invert an optimal morph embedding back to the image space), but proceeds to it quite differently. Moreover the resulting morphing attack do not show very strong success rates compared to the state-of-the-art. We discuss in section 4.2 in what ways our method differs from theirs.

2.2. Template inversion

Several methods have been proposed in the literature to reconstruct face images from facial templates (embeddings) as template inversion attacks against face recognition systems [34, 7, 21, 13, 11, 29, 12, 26, 3, 2]. These methods can be categorized based on the available knowledge from the face feature extractor model into *white-box* and *black-box* methods. In the *white-box* methods, such as [34, 26], the internal functioning and all the parameters of the face feature extractor model are known, and therefore

Table 1. Template Inversion methods in the literature.

Ref.	Reconstruction Quality	Reconstruction Resolution	White-box/ Black-box	Available source code
[34]	low	low	white-box	✗
[7]	low	low	both	✗
[21]	low	low	black-box	✓
[13]	low	low	both	✗
[26]	low	low	white-box	✓
[3]	high	low	black-box	✗
[2]	low	low	black-box	✗
[11]	high	high	black-box	✓
[29]	high	high	black-box	✓
[12]	high	high	black-box	✗
[27]	high	high	both	✓

the feature extractor model is used during training of the face reconstruction network or in gradient-based optimization to reconstruct face images. In contrast, in the *black-box* methods, such as [21, 13, 29], the internal functioning of the face feature extractor model is unknown. Therefore, the feature extractor model cannot be used in the training process of the face reconstruction network, but can be used in non-gradient-based optimizations. Since in the *white-box* methods more knowledge of the feature extractor model is available, it is expected (and shown e.g., in [27]) to achieve better reconstruction performance than *black-box* methods. While most methods are proposed only for either *white-box* or *black-box* scenarios, few methods can be applied to both *white-box* and *black-box* template inversion [7, 13, 27].

Template inversion methods can be also categorized by their output based on the resolution and the quality of reconstruction. Methods that are based on convolutional neural networks (CNNs), such as [34, 26], often generate images that suffer from blurriness or other artifacts. Whereas, most GAN-based methods generate high-quality and realistic (i.e., *human-face-like*) images. In [13], a network based on Pro-GAN [17] is trained with bijection learning to generate realistic face images. Several other methods use StyleGAN to reconstruct face images from facial templates. For instance, in [11, 27] StyleGAN is used as the face generator network and the facial templates are mapped to StyleGAN's first or middle layer. Some other works [29, 12] also used optimization on the input of the StyleGAN to find the latent code that can reconstruct the face image. While the StyleGAN-based methods inherit the leverage of *high-resolution* face generation of StyleGAN, other methods in the literature generate *low-resolution* face images. Table 1 summarizes the template inversion methods proposed in the literature.

3. Methodology

3.1. Morph generation

We introduce a novel method for creating deep morphing attacks, which is grounded in the concept of optimal morph embedding [20]. The process is illustrated in Figure 1.

Given two source images $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$, a facial feature extractor $F(.)$ which extract face embeddings $\boldsymbol{x}_i := F(\boldsymbol{I}) \in \mathcal{X}$, and a distance metric $d(.,.)$ on $\mathcal{X}$, the optimal morph embedding is defined by:

$$\boldsymbol{x}^* := \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left[d(\boldsymbol{x}_1, \boldsymbol{x}) + d(\boldsymbol{x}_2, \boldsymbol{x})\right]. \tag{1}$$

In words, the optimal morph embedding is the face embedding whose biometric distance to the embeddings of both source images is minimized. If in particular the cosine distance is used as metric, and source embeddings are assumed to be normalized, we have

$$\boldsymbol{x}^* := \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{||\boldsymbol{x}_1 + \boldsymbol{x}_2||} \tag{2}$$

An ideal morphing algorithm would only produce face images whose embedding (for each pair of source images) exactly matches the optimal one. However, while the optimal embedding can be computed, it is in principle only a theoretical construct, and transforming it back into the image space is a priori non trivial. Nevertheless, given recent progress in template inversion methods, we believe this transformation is actually feasible and that it will generate a good approximation of optimal morph images.

Our main idea is thus to leverage a biometric template inversion method which will be fed with optimal morph embeddings. With $\mathcal{I}(.)$, a template inversion model trained to invert $\mathcal{F}$, we compute the morph image $\boldsymbol{I}_{morph}$ from the optimal embedding as follows:

$$\boldsymbol{I}_{morph} := \mathcal{I}(\boldsymbol{x}^*) \tag{3}$$

We hypothesize that the resulting images are strong candidates for highly effective morphing. We experiment with two different template inversion systems. The first one (**base inversion**) consists of a self-contained decoder going from the face embedding space back to the image space, which is expected to be very accurate but also produce images of limited quality and resolution (which is illustrated in section 4.1). We thus also experiment with a second inversion system (**GAN-inversion**) which instead learns a mapping from the face embedding space into the latent space of a pretrained StyleGAN model. In doing so, we can leverage the high resolution and realism of StyleGAN generated images, at the possible cost of a lower inversion accuracy. We argue that both those approaches can have their merit depending on whether the *main* focus is to fool the FRS, or to fool some human operator, which is why we choose

to experiment with both methods. The template inversion methods are described in more details in the following section.

3.1.1 Template inversion

To reconstruct the morph images from the optimal morph embeddings, we use state-of-the-art white-box template inversion methods proposed in [26] (for low-resolution morph generation) and in [27] (for high-resolution morph generation). Using a white-box template inversion is particularly desirable in our problem of morph generation because we initially have two face images and extract their embeddings with a feature extractor model. Therefore, it is reasonable to consider a *white-box* template inversion method and use a feature extractor that we have *white-box* knowledge of.

To train the template inversion models, as a preprocessing step, we first normalize the facial templates to have them lie on the embeddings hypersphere (as in Eq. 2), and then train the template inversion network. As mentioned in Section 2.2, the method in [26] generates 112×112 *low-resolution* face images in a *white-box* template inversion and the method in [27] generates 1024×1024 *high-resolution* and *realistic* face images[1]. However, the generated face images by the method proposed in [26] better preserve the identity and achieve a higher attack success rate than the generated face images by the method proposed in [27] in the reported vulnerability evaluation of the same face recognition systems against template inversion attacks.

To train the high-resolution template inversion method based on [27], we use the exact same GAN training proposed in the original work. However, to train the low-resolution template inversion method based on [26], we update the original method to improve the reconstruction quality. Firstly, we applying an additional perceptual loss function :

$$\mathcal{L}_{\text{Perc}}(\hat{\mathbf{I}}, \mathbf{I}) = ||P(\hat{\mathbf{I}}) - P(\mathbf{I})||_1, \tag{4}$$

where $\mathbf{I}$ and $\hat{\mathbf{I}}$ are the original and reconstructed face images, respectively, and P denotes a pre-trained VGG-16 [28] network. Secondly, we also add a skip connection on the convolution blocks.

3.2. Vulnerability evaluation

To study the effectiveness of our approach, we simulate morphing attacks on several FRS and evaluate the attack success rate. We compare it to previous state-of-the-art methods for deep morph generation, mainly StyleGAN interpolation in both the $\mathcal{W}$ space (**SG-W**, as in [24]) and in the $\mathcal{W}+$ space (**SG-W+**, as in [30]), as well as the **MIPGAN** method [32]. We regenerate StyleGAN interpolation morphs using publicly available tools[2]. For MIPGAN morphs, we reuse the code of the original papers that has gracefully been shared with us by the authors.

The evaluation is decomposed in the generation of morphing attacks from a list of images pairs from a source dataset, followed by the actual vulnerability study where the morphing attacks are enrolled into a biometric system then compared against bona fide probes of the contributing subjects. Following the FRONTEX guideline [15], we calibrate the operating threshold to achieve a FMR of 0.1% on a reference bona fide protocol. We then run the vulnerability protocol evaluation (protocol where the morph are enrolled in the system) and report the Mated Morph Presentation Match Rate as introduced in [25], specifically the MinMax-MMPMR and ProdAvg-MMPMR generalizations, at the operating threshold.

We experiment with two sources datasets commonly used in morphing literature, the Face Research Lab London dataset (FRLL) [10] and the Face Recognition Grand Challenge (FRGC) [23]. For FRLL, we select the same morphing pairs as in the AMSL dataset (a morphing dataset also based on FRLL and the morphing method introduced in [22]). For the vulnerability evaluation, we probe the system with all available frontal poses of the contributing subjects. When working with FRGC, we reuse both the morphing pairs and the probes from [32]. We note in particular that as part of developing the MIPGAN system, a StyleGAN instance trained on FFHQ ([18]) has to be fine-tuned on the dataset of contributing subjects. We reuse a MIPGAN system that has been geared towards FRGC morphs, meaning we are not able to also generate FRLL morphs with it. The MIPGAN method will thus only appear in our vulnerability study that uses FRGC as the source dataset.

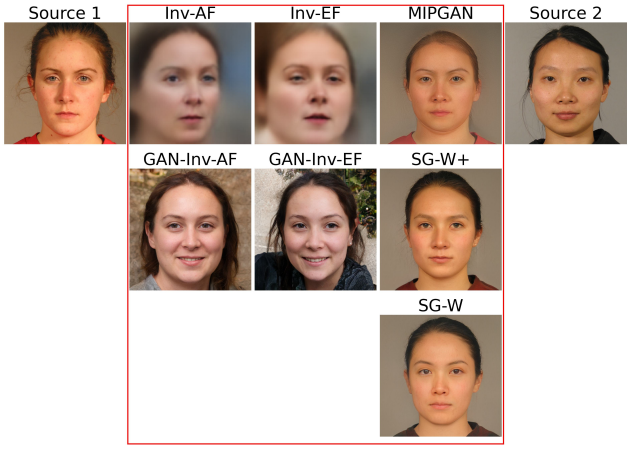
We consider two face recognition systems (FRS), ArcFace [1] and ElasticFace [5]. For each of them, we train our two considered white-box template inversion systems (base inversion and GAN-inversion), resulting in 4 different template inversion systems. For this stage, the FFHQ dataset [18] is used as training data. The same two FRS are then also used as attack target for the vulnerability evaluation. We note that this enables to evaluate how a inversion-based morphing using an inverter trained on some will perform on a different FRS.

The considered morphing approaches are summarized in Table 2. Source code for the replication of those experiments is publicly available.[3]

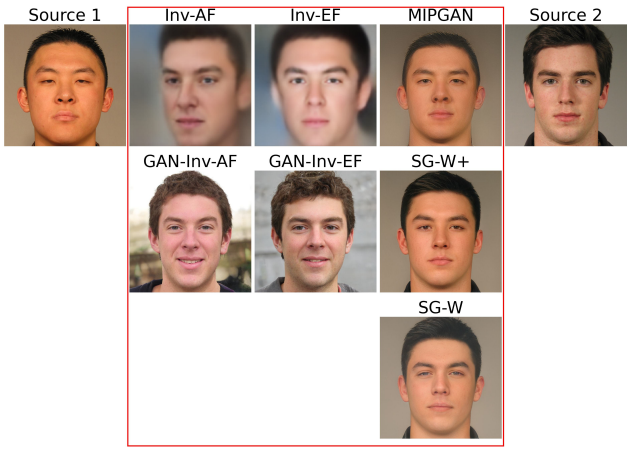
[1] It worth mentioning that the template inversion method proposed in the [27] is the only method that can generate *high-resolution* face images in *white-box* scenario.

Table 2. Considered morphing methods. The inversion and GAN-inversion methods are relative to a specific face recognition system, which will be denoted by either AF (ArcFace) or EF (ElasticFace)

Name	Approach
Inv-AF/EF	Base inversion of optimal template (FRS-dependent)
GAN-Inv-AF/EF	GAN-inversion ($\mathcal{W}$) of optimal template (FRS-dependent)
SG-W[24]	Proj. and interp.. in StyleGAN2 $\mathcal{W}$ space
SG-W+ [30]	Proj. and interp. in StyleGAN2 $\mathcal{W}+$ space
MIPGAN[32]	Optimization in StyleGAN2 $\mathcal{W}+$ space



(a) First pair of sources



(b) Second pair of sources

Figure 2. All types of considered deep morphs for two different pairs of source identities.

4. Results and discussion

4.1. Qualitative discussion

We start with a discussion of the visual aspects of the obtained morphs which are showcased in Figure 2. Optimization-based approaches (SG-W, SG-W+ and MIPGAN) typically produce morphs that look realistic, and whose facial features convincingly seem to be mixing elements from both source faces. We note that SG-W morphs are typically less perceptually similar to the sources than SG-W+ and MIPGAN ones. We also note that MIPGAN morphs regularly showcase blurriness artifacts at the border of the head.

Inversion-based approaches, in contrast, generate morphs that look characteristically more distant from the sources. This is not surprising : the inversion model only has access to face embedding data, which ideally should encapsulate solely information crucial for identification, but no other image features. Non-facial properties (e.g., the background) as well as irrelevant face covariates (pose, expression, hairstyle, etc.) can thus be reconstructed in many different ways and should not be constrained by the content of the source images.

The base inversion method generate morphs that are very blurry outside of the face area, and overall have the worst visual realism of all considered methods. One could argue that those inversion-based morphs are unrealistic to use in a real-world scenario. Indeed, the morph having to be submitted for registration to some passport-issuing authority, it is likely that the image might get processed by a human operator at some point in the process. This operator might easily notice that those morphs are blurry or do not look like a passport photo. While we agree with this remark, we suggest that this problem could be circumvented by further post-processing the resulting morphs to splice the face area (which looks realistic) into one of the source images. This is considered for future work.

In contrast, the GAN-inversion morphs have very high realism, given that they can leverage the richness of StyleGAN's face image distribution modelling. In particular, they do not present easily noticeable artifacts, in contrast to MIPGAN morphs in which some ghostly areas can sometimes be observed around the hair or close to the border of the face. In scenarios where fooling a human operator is crucial for the success of the attack, GAN-inversion morphs might thus be already effective with little to no post-processing.

[2] https://gitlab.idiap.ch/bob/bob.morph.sg2

[3] https://gitlab.idiap.ch/bob/bob.paper.ijcb2023_inversion_morphing

4.2. Quantitative discussion

Tables 3 and 4 present the vulnerability evaluation results, with an operating threshold picked on the *Experiment 2* protocol of the FRGC dataset [23].

We observe that both inversion methods have a significant attack success rate. In particular, when the evaluated FRS is the same that has been inverted to generate the morph, we observe MMPMR in the 40%-50% range for GAN-inversion, and even above 90% for the base inversion method. However, this corresponds to an unrealistic scenario where the attacked network is fully accessible to generate the attack (*white-box attack*[4]). In a real-world setting, it is likely that the attacked FRS would be private, and one could fear that inverted morphs obtained by targeting a specific FRS would only be effective against this same FRS. However we observe that this is not the case : when the evaluated FRS is different from the inverted one (*black-box attack*), we do observe a general decrease in the attack success rate, but it still remains at a concerning level. We also observe that there is some trade-off between realism and success rate : while GAN-inversion morphs are of very high quality, they do not perform as well overall as base inversion ones. In some sense, base inversion could be seen as a method that is biased towards fooling the FRS, at the cost of not fooling humans as much, while GAN-inversion is biased in the opposite direction. We argue that depending on the specific details of the enrolment process (in particular how much it is automated or human-processed), both type of attacks might be relevant.

If we compare inversion-based approaches to previous methods, we observe that our base inversion method is competitive with the state of the art. Indeed, starting with the *white-box* attack scenario, we observe that the Inv-AF systems beats MIPGAN to attack the ArcFace model by a significant margin. We note that MIPGAN should indeed be considered as a white-box attack on ArcFace, given that it uses ArcFace to compute a biometric loss during its own morph generation process. In the *black-box* attack scenario, we observe that the Inv-EF system, in particular, showcases an impressive generalization capability. Indeed, when attacking the ArcFace model, the Inv-EF actually beats MIPGAN, *despite MIPGAN having access to the ArcFace system at morph generation time*. The Inv-AF system also showcases strong black-box attack effectiveness, although it is still beaten by MIPGAN when attacking ElasticFace, but only by a very small margin. Moreover, both base inversion systems always beat SG-W and SG-W+ approaches in the black-box attack scenario, by a wide margin.

[4]We emphasize the distinction between white-box attacks (the attacked FR network is unknown at morph generation time) and white-box template inversion (the inverted FR network is fully accessible to train the inverter). We are here using only white-box inversion methods, but applying them for both white- and black-box morphing attacks.

Table 3. MMPMR on the FRGC vulnerability protocol. Threshold is set for FMR@1e-3 on the FRGC Experiment 2 protocol. The FRS column indicates which face recognition system is used at *evaluation* time. We distinguish between white-box (□) and black-box (■) attacks.

FRS	Attack	MinMax-MMPMR (%)	ProdAvg-MMPMR (%)
AF	□ MIPGAN	73.22	54.77
	□ Inv-AF	**89.88**	**74.76**
	□ GAN-Inv-AF	42.76	22.81
	■ SG-W	4.32	1.44
	■ SG-W+	60.10	39.97
	■ Inv-EF	**79.65**	**61.46**
	■ GAN-Inv-EF	16.46	5.85
EF	□ Inv-EF	**87.78**	**74.58**
	□ GAN-Inv-EF	28.88	14.52
	■ SG-W	10.19	3.56
	■ SG-W+	67.63	48.90
	■ MIPGAN	**75.80**	**60.10**
	■ Inv-AF	75.09	58.25
	■ GAN-Inv-AF	28.20	14.73

The GAN-inversion method, however, is not as effective as previous morphing methods, even though it still showcases concerning attack success rates. But we want to emphasize that this method is learning a mapping from the face embedding space into the $\mathcal{W}$ space of the used StyleGAN network, which has a limited capacity. As illustration, we see for example that switching from $\mathcal{W}$ to $\mathcal{W}+$ with the StyleGAN morphing methods drastically improves the attack success rate. MIPGAN is also finding the morph by exploring the $\mathcal{W}+$ space. We hypothesize that the performance of our GAN-inversion system might be partially limited by this restriction to the $\mathcal{W}$ space, and that learning a new encoder of face embeddings into the $\mathcal{W}+$ space might give significant returns. However, the process is trickier to train given the high dimensionality of the output space. This is left for future work.

We also want to discuss how our method compared by the one proposed in [20]. This work also attempts to invert an optimal morph embedding by (we stick to the formalism of section 3.1)

1. complementing the facial feature extractor with a second encoder E trained to extract image features *not* related to the identity,
2. training a decoder D to reconstruct an image from $E(\boldsymbol{I}_1)$ and $\boldsymbol{x}^*$, which should perceptually look like $\boldsymbol{I}_1$, but have a face embedding close to $\boldsymbol{x}^*$.

In some sense, this process learns to introduce *imperceptible* signal guided by $\boldsymbol{I}_2$ onto $\boldsymbol{I}_1$, to bring its face embedding much closer to $\boldsymbol{x}^*$. While effective, this method's main drawback is its reliance on the image features encoder E.

Table 4. MMPMR on the FRLL vulnerability protocol. Threshold is set for FMR@1e-3 on the FRGC Experiment 2 protocol. The FRS column indicates which face recognition system is used at *evaluation* time. We distinguish between white-box (□) and black-box (■) attacks.

FRS	Attack	MinMax-MMPMR (%)	ProdAvg-MMPMR (%)
AF	□ Inv-AF	**97.54**	**94.47**
	□ GAN-Inv-AF	51.58	42.39
	■ SG-W	1.05	0.64
	■ SG-W+	62.63	53.71
	■ Inv-EF	**90.70**	**85.57**
	■ GAN-Inv-EF	17.19	11.93
EF	□ Inv-EF	**96.67**	**93.00**
	□ GAN-Inv-EF	37.46	29.10
	■ SG-W	3.07	2.06
	■ SG-W+	72.28	62.81
	■ Inv-AF	**91.75**	**86.80**
	■ GAN-Inv-AF	41.93	33.07

As this encoder in particular learns general properties of the image distribution of the training set (e.g. color or textures), it might struggle with generalizing to other source datasets (which is unfortunately not evaluated in the original paper), which could show a different color or textures distribution. In contrast, we showcased the reliance on this image features encoder is actually superfluous, and that effective inverted morphs can be obtained solely using optimal morph embeddings as input.

Finally, inversion-based morphing has additional advantages on top of its high attack effectiveness. Firstly, previous methods rely on a time consuming optimization process exploring the latent space of StyleGAN in order to find either a good projection of the source images (SG-W, SG-W+) or directly a candidate latent that generates an effective morph (MIPGAN). In contrast, once the inversion model is trained, generating inverted morphs is a straightforward process that only requires two forwards passes of the face recognition network and one forward pass of the template inverter. For this reason, morphs can be generated at a speed orders of magnitude faster. This is showcased in Table 5 which presents typical runtimes for end-to-end generation of a single morph. We observe that any of the inversion-based approaches leads to a speed up of around 50x - 75x with respect to MIPGAN. This major speed up could greatly facilitate the creation of large deep morph datasets; for example, to enable the training of effective detection models. Secondly, the MIPGAN model is relying on a fine-tuning of a pretrained StyleGAN-FFHQ model using the source dataset. It is yet unclear how much the resulting morphing system is sensitive to the similarity of the source images to the fine-tuning dataset, and it might not always be simple to assemble a new adapted fine-tuning dataset to recalibrate the system for source images that are out-of-distribution. We showcased that inversion-based morphing does not suffer from such limitation, as it only relies on the FFHQ dataset at training, but can then be used out-of-the-box on various source data (here FRLL and FRGC) while showing similar success.

Table 5. End-to-end runtime of each generation algorithm to create 1 morph. The measurements are averaged over 10 morph generations.

Attack	Runtime [s]
SG-W	372.08 ± 1.46
SG-W+	373.67 ± 2.77
MIPGAN	47.43 ± 1.64
Inv-AF	0.88 ± 0.05
Inv-EF	0.64 ± 0.01
GAN-Inv-AF	0.99 ± 0.05
GAN-Inv-EF	0.75 ± 0.01

4.3. Remark on the detectability of inversion-based morphs

As we introduce this new deep morphing attack generation method which shows high effectiveness, we are concerned whether it could also challenge common morphing attack detection systems. We believe that the new proposed method should *not* be drastically more difficult to detect as previous ones. GAN-inversion morphs in particular are still in the end a particular output of a StyleGAN model. StyleGAN images can be reliably detected as showcased in [31], even when not sampled in a straightforward manner : [6] showcases that SG-W+ and MIPGAN morphs can also be reliably detected. For morphs obtained with our base inversion method, we do not have as many guarantees; however [31] claims that the main salient fingerprint of image generative models is caused by upsampling artifacts in the convolutional architecture, a signal that should thus also be present in our inverted morphs given the architecture of the inverter. To verify this, we propose to actually run our morphs through the detection model from [31]. It is a GAN-image detector that showcases strong generalization to unknown generators. We run through this detector morphs sets derived from FRGC with all of our considered morphing methods, from which we get a distribution of attack scores. We use a subset of 1000 images of the FFHQ dataset [18] to generate a set of bonafide scores. We then report in table 6 the AUC as well as the equal error rate of the detector using those two sets as respectively positive and negative examples. This corroborates our hypothesis that both base inversion and GAN-inversion morphs can still be detected with reasonable accuracy, but still not as well as previous methods. Improvements in the robustness of existing detectors to new types of morphing attacks is thus still needed. Works in this line of research would benefit from

Table 6. Performance of the detector from [31] on morphing sets derived from FRGC. 1000 images of FFHQ form the bonafide set.

Attack	AUC	EER (%)
Inv-AF	0.974	7.30
Inv-EF	0.948	11.61
GAN-Inv-AF	0.977	6.70
GAN-Inv-EF	0.982	5.62
SG-W+	1.000	0.80
MIPGAN	1.000	0.18

having access to deep morphs datasets showcasing a wide variety of attacks, however such datasets are still scarcely available. We hope our work can contribute to mitigating this scarcity.

5. Conclusion

We have demonstrated the feasibility of generating morphing attacks by leveraging template inversion systems to invert optimal morph embeddings. Both our methods significantly improve the generation speed with respect to the previous state of the art. Moreover, our base inversion morphing method is competitive with the previous state-of-the-art in terms of attack success rate, and often beats it by a large margin, both in the white-box and black-box attack scenario. The Inv-EF system in particular showcases such strong generalization that even in a black-box attack scenario (when attacking the ArcFace model), it is still more effective that MIPGAN, despite the latter actually using ArcFace to compute a biometric loss as part of the morph generation process.

One main limitation of this base inversion method is that the resulting morphs are somewhat lacking in realism. We believe that a further post-processing of those morphs to splice them back into one of the source images could mitigate the visual realism issue while not losing too much in attack success rate. Our GAN-inversion morphing method does display great realism, but generates attacks with lower (but still problematic) effectiveness. We hypothesize that this latter method could be improved by mapping face embeddings into the $\mathcal{W}+$ space of StyleGAN (which has greater capacity) instead of the $\mathcal{W}$ space as is done currently. Indeed, we note that with StyleGAN interpolation methods for example, the simple switch from $\mathcal{W}$ to $\mathcal{W}+$ drastically improve the effectiveness of the attack. Interestingly, despite their reliance on a white-box access to some FRS, the inversion morphing attacks stays successful when used to attack some other unseen FRS. There is still however a decrease in effectiveness in this latter case; improving the generalization to unseen FRS is another direction that might be interesting for future work.

Finally, our methods enable fast generation of large scale morphing datasets, which we hope could facilitate the development and training of deep morphing attack detection systems.

Acknowledgments

This research is based upon work supported by the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813). This work was also supported by the Swiss Center for Biometrics Research & Testing and the Idiap Research Institute.

References

[1] ArcFace: Additive Angular Margin Loss for Deep Face Recognition. https://ieeexplore.ieee.org/document/8953658. 4

[2] S. Ahmad, K. Mahmood, and B. Fuller. Inverting biometric models with fewer samples: Incorporating the output of multiple models. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022. 2, 3

[3] M. Akasaka, S. Maeda, Y. Sato, M. Nishigaki, and T. Ohki. Model-free template reconstruction attack with feature converter. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022. 2, 3

[4] Z. Blasingame and C. Liu. Leveraging Diffusion For Strong and High Quality Face Morphing Attacks, Feb. 2023. 2

[5] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper. ElasticFace: Elastic Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2022. 4

[6] L. Colbois and S. Marcel. On the detection of morphing attacks generated by GANs. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2022. 7

[7] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3703–3712, 2017. 2, 3

[8] N. Damer, K. Raja, M. Süßmilch, S. Venkatesh, F. Boutros, M. Fang, F. Kirchbuchner, R. Ramachandra, and A. Kuijper. ReGenMorph: Visibly Realistic GAN Generated Face Morphing Attacks by Attack Re-generation. In G. Bebis, V. Athitsos, T. Yan, M. Lau, F. Li, C. Shi, X. Yuan, C. Mousas, and G. Bruder, editors, *Advances in Visual Computing*, Lecture Notes in Computer Science, pages 251–264, Cham, 2021. Springer International Publishing. 2

[9] 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 (BTAS)*, pages 1–10, Oct. 2018. 2

[10] L. DeBruine and B. Jones. Face Research Lab London Set, May 2017. 4

[11] X. Dong, Z. Jin, Z. Guo, and A. B. J. Teoh. Towards generating high definition face images from deep templates. In *Proceedings of the International Conference of the Biometrics*

Special Interest Group (BIOSIG), pages 1–11. IEEE, 2021. 2, 3

[12] X. Dong, Z. Miao, L. Ma, J. Shen, Z. Jin, Z. Guo, and A. B. J. Teoh. Reconstruct face from features using gan generator as a distribution constraint. *arXiv preprint arXiv:2206.04295*, 2022. 2, 3

[13] C. N. Duong, T.-D. Truong, K. Luu, K. G. Quach, H. Bui, and K. Roy. Vec2face: Unveil human faces from their blackbox features in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6132–6141, 2020. 2, 3

[14] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, Sept. 2014. 2

[15] FRONTEX. Best practice technical guidelines for automated border control ABC systems, 2015. 4

[16] J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2

[17] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 3

[18] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, June 2019. 4, 7

[19] 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*, pages 8110–8119, 2020. 2

[20] U. M. Kelly, L. Spreeuwers, and R. Veldhuis. Worst-Case Morphs: A Theoretical and a Practical Approach. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, Sept. 2022. 2, 3, 6

[21] G. Mai, K. Cao, P. C. Yuen, and A. K. Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2018. 2, 3

[22] 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.*, 2018. 4

[23] 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)*, volume 1, pages 947–954 vol. 1, June 2005. 4, 6

[24] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Are GAN-based morphs threatening face recognition? In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2959–2963, May 2022. 2, 4, 5

[25] U. Scherhag, A. Nautsch, C. Rathgeb, M. Gomez-Barrero, R. N. J. Veldhuis, L. Spreeuwers, M. Schils, D. Maltoni, P. Grother, S. Marcel, R. Breithaupt, R. Ramachandra, and C. Busch. Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting. In *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–7, Sept. 2017. 4

[26] H. O. Shahreza, V. K. Hahn, and S. Marcel. Face reconstruction from deep facial embeddings using a convolutional neural network. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1211–1215. IEEE, 2022. 1, 2, 3, 4

[27] H. O. Shahreza and S. Marcel. Face reconstruction from facial templates by learning latent space of a generator network. In https://openreview.net/pdf?id=j1HyTEWHTT, 2023. 1, 3, 4

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[29] E. Vendrow and J. Vendrow. Realistic face reconstruction from deep embeddings. In *Proceedings of NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. 2, 3

[30] 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)*, pages 1–6, Apr. 2020. 2, 4, 5

[31] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 7

[32] H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch. MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):365–383, July 2021. 2, 4, 5

[33] N. Zhang, X. Liu, X. Li, and G.-J. Qi. MorphGANFormer: Transformer-based Face Morphing and De-Morphing, Feb. 2023. 2

[34] A. Zhmoginov and M. Sandler. Inverting face embeddings with convolutional neural networks. *arXiv preprint arXiv:1606.04189*, 2016. 2, 3