Master Face Attacks on Face Recognition Systems

Huy H. Nguyen, *Member, IEEE*, Sébastien Marcel, *Senior Member, IEEE*, Junichi Yamagishi, *Senior Member, IEEE*, and Isao Echizen, *Member, IEEE*

***Abstract*—Face authentication is now widely used, especially on mobile devices, rather than authentication using a personal identification number or an unlock pattern, due to its convenience. It has thus become a tempting target for attackers using a presentation attack. Traditional presentation attacks use facial images or videos of the victim. Previous work has proven the existence of master faces, i.e., faces that match multiple enrolled templates in face recognition systems, and their existence extends the ability of presentation attacks. In this paper, we report an extensive study on latent variable evolution (LVE), a method commonly used to generate master faces. An LVE algorithm was run under various scenarios and with more than one database and/or face recognition system to identify the properties of master faces and to clarify under which conditions strong master faces can be generated. On the basis of analysis, we hypothesize that master faces originate in dense areas in the embedding spaces of face recognition systems. Last but not least, simulated presentation attacks using generated master faces generally preserved the false matching ability of their original digital forms, thus demonstrating that the existence of master faces poses an actual threat.**

***Index Terms*—Master face, wolf attack, face recognition system, latent variable evolution.**

I. Introduction

PASSWORDS should be strong, which can make them difficult to remember, and should be changed regularly to ensure security. Personal identification numbers and unlock patterns are more convenient than passwords, but the user is still required to remember them, and people nearby may be able to steal a peek at them. An even more convenient method is biometric authentication, which uses a biometric trait unique to the user, eliminating the need to remember anything. This advantage has led to the widespread usage of biometric authentication on many portable devices including laptops and smartphones. The two most commonly used biometric traits for authentication are a fingerprint and the face [1]. Since smartphones using this type of authentication may have a digital wallet (or e-wallet) for making e-payments, they are a prime target for attackers. An attacker may attempt to unlock such a device by performing a presentation attack [2]. For example, the attacker might attempt a presentation attack in which a printed facial image of the victim (known as a *presentation attack instrument, or PAI*) is displayed in front of the smartphone's camera.

The probability of a presentation attack succeeding is higher if the PAI matches multiple enrolled templates. In the facial domain, the creation of PAIs by blending together two or more faces is called *face morphing* [5]. The morphed face should match all source faces when used against a face recognition (FR) system and possibly even fool a human observer. This ability has made morphing a commonly used attack against automated border control systems in which the attacker "borrows" the identity of the victim to enter or exit a location [5]. The face morphing approach is limited by the requirement that target faces be available. Another approach is to generate a "master biometric" sample [4], [6]–a kind of "wolf sample" that matches multiple enrolled templates in a biometric recognition system [7]. This approach was first developed by Bontrager *et al.* [6] for the fingerprint domain. In our previous work [4] and this extended work, we have adopted this approach and extended it to the facial domain. Unlike the face morphing approach, the attacker's advantage in this "master face" approach is that it does not require any information about the victim. Moreover, using an ordinary PC and materials easily obtained from the Internet is enough to generate master faces. Before this work, the nature and characteristics of master faces were not well (or sufficiently) understood.

The stages in master biometrics research are shown in Fig. 2. Our contributions can be summarized as follows:

- Building on our previous work [4], we are the first to generate master faces that can match multiple faces with different identities. This ability means that FR systems are vulnerable to a master face attack.
- We extend our previous work by analyzing the effect of using multiple databases (DBs) and/or multiple FR systems for the latent variable evolution (LVE) algorithm used to generate master faces. Some DB/FR system combinations boosted overall attack performance while others did not due

Manuscript received 9 July 2021; revised 9 January 2022; accepted 6 April 2022. Date of publication 15 April 2022; date of current version 15 July 2022. This work was supported in part by JSPS KAKENHI under Grant JP16H06302, Grant JP18H04120, Grant JP21H04907, Grant JP20K23355, and Grant JP21K18023; and in part by JST CREST under Grant JPMJCR18A6 and Grant JPMJCR20D3, including the AIP Challenge Program, Japan. This article was recommended for publication by Associate Editor R. Singh upon evaluation of the reviewers' comments. *(Corresponding author: Huy H. Nguyen.)*

Huy H. Nguyen was with the Graduate University for Advanced Studies, SOKENDAI, Kanagawa 240-0193, Japan. He is now with the National Institute of Informatics, Tokyo 101-8430, Japan (e-mail: nhhuy@nii.ac.jp).

Sébastien Marcel is with the Biometrics Security and Privacy Group, Idiap Research Institute, 1920 Martigny, Switzerland.

Junichi Yamagishi is with the National Institute of Informatics, Tokyo 101-8430, Japan, and also with the Graduate University for Advanced Studies, SOKENDAI, Kanagawa 240-0193, Japan.

Isao Echizen is with the National Institute of Informatics, Tokyo 101-8430, Japan, also with the Graduate University for Advanced Studies, SOKENDAI, Kanagawa 240-0193, Japan, and also with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113-8654, Japan.

Digital Object Identifier 10.1109/TBIOM.2022.3166206

Fig. 1. Original master face generated using two face recognition systems (top left) and its PAI forms printed on plain paper (top right), photo paper (bottom left), and displayed on a 13-inch Apple MacBook Pro screen (bottom right).

to intra-component conflicts. Knowledge of the successful combinations is critical to understanding under which conditions strong master faces can be generated and to appropriately assessing the potential risks.

- We expand the scope of our previous work by introducing more scenarios and by using an additional facial database and an additional FR system trained with the angular margin loss [8] for the defender side. Furthermore, we introduce visualization in the face embedding (identity) space to obtain more insights into master faces. These insights are invaluable as they can be used to improve the robustness of FR systems.
- To demonstrate the actual threat posed by the existence of master faces, we evaluated master face attacks by performing presentation attacks using printed images and the corresponding digital images displayed on a computer screen. Three of the PAIs we used are shown in Fig. 1.

The rest of the paper is organized as follows. First, we provide background information on facial image generation, FR systems, wolf attacks, master biometric attacks, and the LVE algorithm in Section II. Then, we discuss the existence of master faces and introduce an improved LVE algorithm using multiple databases and/or FR systems in Section III. Our experiments are covered in two sections: we first discuss generating master faces and their analysis in Section IV and then discuss using master faces to perform presentation attacks in Section V. Next, in Section VI, we discuss ways to reduce the risk of master face attacks. Finally, we summarize the key points and make some closing remarks in Section VII.

II. Related Work

A. Facial Image Generation

Image generation is a major topic in deep learning research, and the face is a common target. There are two major approaches to image generation: using variational autoencoders (VAEs) [9] and using generative adversarial networks (GANs) [10]. In the beginning, they could only generate small images with low quality. VAEs tended to generate blurry images while GANs were difficult to train. Subsequent improvements in GANs (WGAN [11] and WGAN Gradient Penalty (WGAN-GP) [12]) resolved the training problem, and GANs then began to be used to generate master prints [6].

Recently improved versions of both VAEs [13], [14] and GANs [15]–[18] can generate high-resolution images. By gradually adding more layers during training to output larger images, Karras *et al.* were able to generate 1024 $\times$ 1024 pixel images with their progressive GAN [16]. In subsequent work, they combined the ideas of progressive training and style transfer to create a better disentanglement network called *StyleGAN* [17]. Unlike traditional GANs, which directly use a latent vector for generating images, StyleGAN uses a mapping network to transfer this latent vector into intermediate style vectors used for synthesizing images. Controlling these intermediate style vectors changes the facial attributes. With the abilities of strong disentanglement and high-quality facial image generation, StyleGAN and its subsequent version [18] are the best methods for generating master faces [4].

B. Face Recognition

The release of large databases (e.g., the CASIA-WebFace database [19] and the MS-Celeb database [20]) and recent advances in convolutional neural networks (CNNs) have substantially improved the performance of FR systems and enabled them to work effectively in heterogeneous domains [21]. Most state-of-the-art FR systems [8], [21], [22] make use of a network architecture that achieved high performance in the ImageNet Challenge [23], such as the VGG (Visual Geometry Group) network architecture [24] and the inception network architecture [25]. Parkhi *et al.* trained the VGG-16 network on a custom-built large-scale database [26] to create the VGG-Face network. Wu *et al.* proposed a lightweight CNN that has ten times fewer parameters than the VGG-Face network [27]. The inception architecture was used by de Freitas Pereira *et al.* to build heterogeneous FR networks [21] and by Schroff *et al.* to build the FaceNet network [22]. Sandberg re-implemented FaceNet as an open-source system [28]. Taigman *et al.* introduced DeepFace in which explicit 3D face modeling is used to improve the facial alignment phase and a CNN is used to extract face representation [29]. Unlike previous methods, which use discriminative classifiers, the generative classifier proposed by Tran *et al.*, called *DR-GAN*, learns a disentangled representation [30].

More recent approaches focus on optimizing the embedding distribution. Deng *et al.* proposed using the additive angular margin loss (ArcFace) instead of the commonly used cosine distance loss to improve the discriminative power of the FR model and to stabilize the training process [8]. Duan *et al.* argued that the distribution of the features plays an important role and therefore proposed using a uniform loss to learn equidistributed representations for their UniformFace FR system [31].

FR systems are vulnerable to presentation attacks, which present an artifact or human characteristic to the biometric

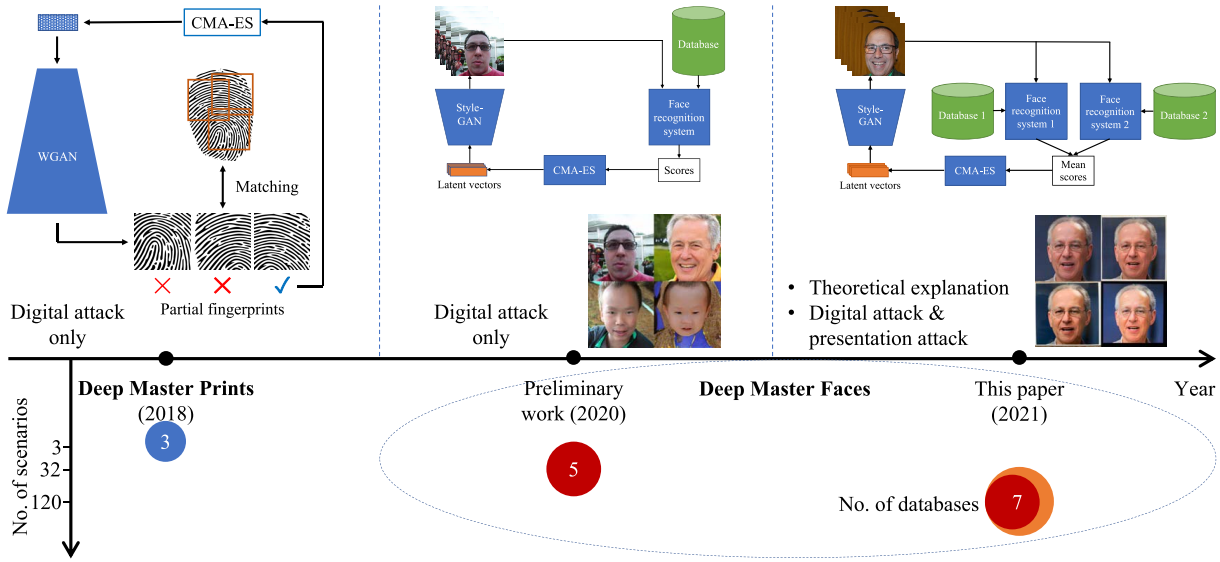

Fig. 2. Stages in master biometrics research. First stage was partial master fingerprints, as proposed by Bontrager *et al.* [3]. Next stage was our preliminary work on master faces [4]. Current stage (this work) builds upon previous work and introduces extensions in algorithm, analysis, visualization, and test scenarios.

(facial) capture subsystem to interfere with the intended policy of the biometric (FR) system.[1] A photo attack is a presentation attack in which the attacker displays a photograph of the victim to the sensor of the FR system. This photograph can be printed on paper or displayed on a device's screen (e.g., a smartphone, a tablet, or a laptop) [32]. A replay attack is another presentation attack in which a victim's video is played instead of displaying a photograph [32]. A presentation attack detector can be integrated into an FR system to mitigate presentation attacks [32].

To study the security threat posed by master faces, we built on some of the advances discussed above and conducted rigorous experiments with four recent (and conceptually diverse) state-of-the-art FR systems. Please note that a presentation attack detector was not integrated into these FR systems.

C. Wolf Attack and Master Biometric Attack

A "wolf sample" is an input sample that can be falsely accepted as a match with multiple user templates ("enrolled subjects") in a biometric recognition system [7]. Wolf samples can be either biometric or non-biometric. A wolf sample is used in a wolf attack against a biometric recognition system. An example wolf attack is shown in Fig. 3. Wolf attacks were initially used to target fingerprint recognition systems [33]. Success is theoretically characterized by the wolf attack probability (WAP)–the maximum probability of a successful attack with one wolf sample [7]. Inuma *et al.* [34] presented a principle for mitigating wolf attacks against biometric authentication systems: construct a secure matching algorithm that calculates the entropy of the probability distribution of each input value.

A master biometric attack is a wolf attack in which the sample looks like an actual biometric trait. Two example traits are partial fingerprint images [6] and facial images [4].

[1]ISO/IEC CD 30107-1 definition. Accessed at [https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en:term:3.5?]

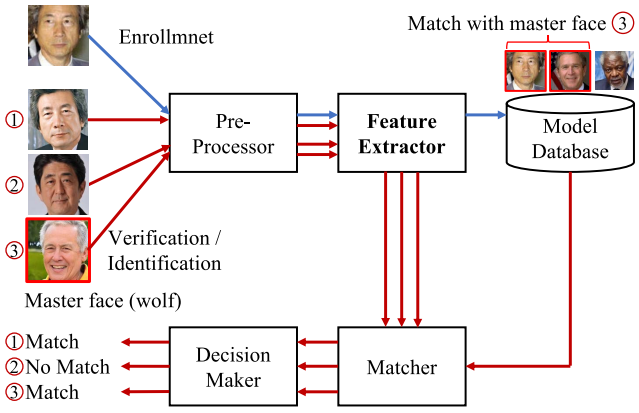

Fig. 3. Operation of typical FR system. There are two phases: enrollment (blue path) and verification/identification (red path). The master face (face 3) was falsely matched with the two faces of two enrolled subjects. Best viewed in color.

They are generated by GANs using the LVE algorithm to maximize the false matching rates (as a result, WAPs are also maximized). A master print attack [6] targets partial fingerprint recognition systems using small sensors with limited resolution while a master face attack targets FR systems, which require higher resolution images [4]. In this work, we used multiple FR systems and databases when running the LVE algorithm to generate master faces. We also simulated presentation attacks using master faces to ascertain their actual threat.

D. Latent Variable Evolution

Evolution algorithms are commonly used in artificial intelligence applications to approximate complex, multimodal, and non-differentiable functions since they do not require any assumption about the underlying fitness landscape. The covariance matrix adaptation evolution strategy (CMA-ES) is a powerful strategy designed for non-linear and non-convex

functions [35]. Bontrager *et al.* used CMA-ES with a pre-trained GAN to perform interactive evolutionary computation to improve the quality of generated samples [3]. This strategy was used in subsequent work on the LVE algorithm to maximize the WAP of generated partial fingerprint images [6]. In our previous work [4], we modified the LVE algorithm scoring method so that it could work smoothly with high-resolution facial images generated by StyleGAN [17].

Given n random initial vectors $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$, a generation model $\mathcal{G}$, a scoring function $\mathcal{F}$, and m enrolled templates $\mathcal{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_m\}$, the LVE algorithm runs in a loop in which n samples are first generated by $\mathcal{G}$ using $\mathcal{Z}$. Each sample is then matched with m templates in $\mathcal{T}$ to obtain a mean score s. An evolution algorithm (e.g., CMA-ES) takes the set of the mean scores s to evolve n new latent vectors $\mathcal{Z}'$ for the next loop.

We have now added one more database and/or FR system to the LVE algorithm to better approximate the target FR system and database so that the generated master faces have better generalizability.

III. Deep Master Faces

A. Existence of Master Faces

Before describing the proposed master face generation algorithm, we briefly explain why master faces exist. For a typical FR system (or biometric recognition systems in general), there are four phases (Fig. 3): pre-processing the input, extracting its features, matching them with those of the enrolled subject(s) in the model database, and making a decision. The feature extractor plays the role of a mapping function. It maps the facial image domain to the identity domain. The objective when training the feature extractor is to optimize the mapping function so that the mappings of the same-identity faces are close together in the identity space and vice versa. Since this is an optimization problem, the solution is simply an approximation. Furthermore, there is no guarantee that the mapping function will work well on new data due to the possible lack of generalizability.

Master faces may exist because the identity (embedding) space used by FR systems is not uniformly distributed, resulting in dense areas in this space. If we generate an identity corresponding to a point in a dense area, it may falsely match several nearby faces in the identity space. The LVE algorithm aims to find such a position in a dense area in the identity space after several evolutions. To intuitively and empirically show this, we visualize the identity space and one of the master faces generated in this work using uniform manifold approximation and projection (UMAP) [36] in Fig. 4. The master face generated by our algorithm (described in the next section) is at such a position (red dot) surrounded by many embeddings. All faces from these surrounding embeddings are falsely matched with the master face by the FR system. The no-match embeddings are scattered far from the master face and lie in less dense areas. We explain how to generate such master faces in the next section.

To verify our hypothesis of dense areas in the identity space, we searched for the real faces that were closest to a

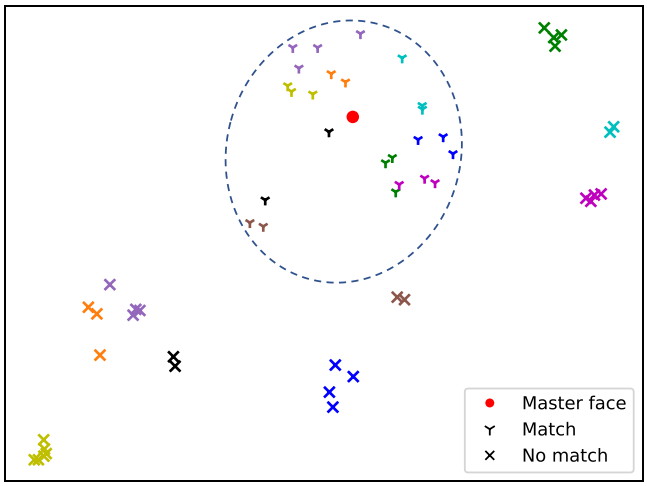

Fig. 4. UMAP visualization of identity space containing embeddings of a master face and of "match" and "no-match" faces of 18 enrolled subjects. For each cluster (match or no match), symbols with the same color correspond to the same subject. Best viewed in color.

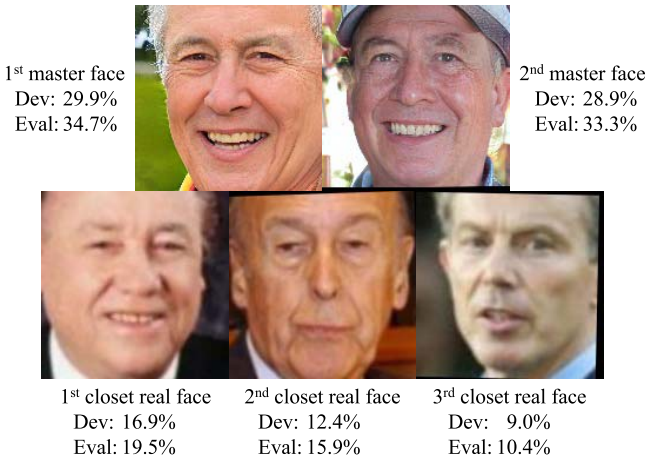

Fig. 5. The first and second master faces generated using the LVE algorithm and the three real faces closest to the first master face and their corresponding false matching rate. The two master faces were generated using the training set of the LFW - Fold 1 database and the Inception-ResNet-v2 based FR system trained on the CASIA-WebFace database. The false matching rates were calculated on the development and evaluation sets of the Labeled Faces in the Wild (LFW) - Fold 1 database.

generated master face and checked whether they had wolf characteristics like this master face. These real faces were chosen from the facial database used to generate the master face. We used the cosine distance between two embeddings for selection. The result, which is shown in Fig. 5, confirms our hypothesis. However, the real wolf faces had lower false matching rates (FMRs) than the master faces. Therefore, using synthesized master faces rather than real faces to carry out wolf attacks should increase the success rate.

B. Latent Variable Evolution With Multiple Databases and/or Face Recognition Systems

We extended our previous work by using one more database and/or FR system to generate master faces, which requires support from the LVE algorithm. The extended LVE algorithm

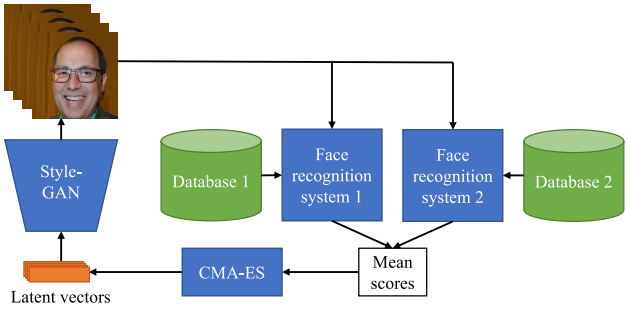

Fig. 6. Overview of extended LVE algorithm. Latent vectors are fed into StyleGAN [17] to generate facial images. One or more surrogate FR system(s) then calculates mean score for each image on the basis of the subjects in one or more database(s). For example, for the *combination 3* setting described in Table III, **database 1** is LFW - Fold 1, **database 2** is mobile biometry (MOBIO), **FR system 1** is Inception-ResNet-v2 network (trained on MS-Celeb database), and **FR system 2** is DR-GAN network. The CMA-ES [35] algorithm uses these scores to generate new latent vectors.

Algorithm 1 Latent Variable Evolution

$m \leftarrow 22$ ▷ Population
procedure RUNLVE(m, n)
  $\mathcal{F} = \{\}$ ▷ Master face set
  $\mathcal{S} = \{\}$ ▷ and corresponding score set
  $\mathcal{Z} = \{\mathbf{z_1} \leftarrow rand(), \ldots, \mathbf{z_m} \leftarrow rand()\}$ ▷ Initialize
  for n iterations **do** ▷ Run LVE algorithm n times
    $F \leftarrow StyleGAN(\mathcal{Z})$ ▷ Generate m faces F
    $\mathbf{s^{(1)}} \leftarrow 0, \mathbf{s^{(2)}} \leftarrow 0$ ▷ Initialize scores $\mathbf{s^{(1)}}, \mathbf{s^{(2)}} \in \mathbb{R}^m$
    for face F_i in faces **F** **do**
      for face $E_j^{(1)}$ in data $\mathbf{E^{(1)}}$ **do**
        $s_i^{(1)} \leftarrow s_i^{(1)} + FaceMatching^{(1)}(F_i, E_j^{(1)})$
      $s_i^{(1)} \leftarrow \frac{s_i^{(1)}}{|\mathbf{E^{(1)}}|}$ ▷ Mean scores of 1st system
      for face $E_j^{(2)}$ in data $\mathbf{E^{(2)}}$ **do**
        $s_i^{(2)} \leftarrow s_i^{(2)} + FaceMatching^{(2)}(F_i, E_j^{(2)})$
      $s_i^{(2)} \leftarrow \frac{s_i^{(2)}}{|\mathbf{E^{(2)}}|}$ ▷ Mean scores of 2nd system
      $s_i = \frac{s_i^{(1)} + s_i^{(2)}}{2}$ ▷ Mean scores of both systems
    $F_b, s_b \leftarrow GetBestFace(\mathbf{F}, \mathbf{s})$
    $\mathcal{F} \leftarrow \mathcal{F} \cup \{F_b\}$ ▷ Append best master face
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_b\}$ ▷ and its corresponding score
    $\mathcal{Z} \leftarrow$ CMA_ES(**s**)
  return $\mathcal{F}, \mathcal{S}$
$F_b, s_b \leftarrow GetBestFace(\mathcal{F}, \mathcal{S})$ ▷ Final (best) master face

is illustrated in Fig. 6 and is formalized in Algorithm 1. First, m latent vectors $\{\mathbf{z_1}, \ldots, \mathbf{z_m}\}$ are initialized randomly. They are then fed into a pretrained StyleGAN network to generate m faces. Two face matching functions, $FaceMatching^{(1)}(\cdot,\cdot)$ and $FaceMatching^{(2)}(\cdot,\cdot)$ (corresponding to two FR systems), calculate the similarity between the generated faces and all subject faces in databases $E_j^{(1)}$ and $E_j^{(2)}$, respectively. Two m-dimension mean score vectors, $\mathbf{s^{(1)}}$ and $\mathbf{s^{(2)}}$, are obtained from the results of $FaceMatching^{(1)}(\cdot,\cdot)$ and $FaceMatching^{(2)}(\cdot,\cdot)$. The mean **s** of these two vectors is used to select the best local master face F_b among the m generated faces. Finally, **s** is fed into the CMA-ES algorithm to generate new latent vectors $\{\mathbf{z_1}, \ldots, \mathbf{z_m}\}$. This process is repeated n times. The final (global) best master face is chosen from among the n best master faces $\mathcal{F}$ obtained in the n iterations.

Algorithm 2 Database Refining

$\mathcal{M} = \{M_1, \ldots, M_n\}$ ▷ Previous master faces
procedure REFINE_DATABASE($\mathcal{M}$, **E**)
  $\mathbf{E}' = \{\}$ ▷ Initialize refined database
  for face E_i in data **E** **do**
    keep $\leftarrow$ **true**
    for face M_j in $\mathcal{M}$ **do**
      if isMatch(E_i, M_j) is **true then**
        keep $\leftarrow$ **false**
    if keep is **true then**
      $\mathbf{E}' \leftarrow \mathbf{E}' \cup \{E_i\}$
  return $\mathbf{E}'$

To generate another master face, all faces matching the previously generated master face(s) in the training database(s) need to be removed, as shown in Algorithm 2. This prevents the new master face from overlapping the previous master face(s). An example of a second master face is shown in Fig. 5 along with the first master face, the real wolf face, and their corresponding FMRs The FMR of the second master face is lower than that of the first one, and this usually holds for any subsequent master faces.

IV. Generating Master Faces

To evaluate the risks and threats of a master face attack, we designed several settings for the LVE algorithm and several attack scenarios that cover white-box, gray-box, and black-box attacks. For white-box attacks, both the architecture of the target FR system and its training database are known while for gray-box attacks, only one of them is known. For black-box attacks, there is no information about the target FR system. Attackers may use more than one FR system for the LVE algorithm to increase the probability of their attack being a white-box or gray-box attack. They can also use more than one database for the LVE algorithm to better approximate the distribution of the model database of the target FR system.

This section is organized as follows: We first briefly describe the FR systems and the databases we used in our experiments. Then, we describe our generation of master faces using several combinations of single and multiple FR systems with single and multiple facial databases when running the LVE algorithm. Next, we analyze the generation processes and the generated master faces as well as explain their properties. Finally, we evaluate the false matching performance of the generated master faces for several scenarios, including black-box, gray-box, and white-box attacks.

A. Experiment Materials

1) Face Recognition Systems: We used five mainstream publicly available high-performance FR systems in our experiments:

- Inception-ResNet-v2 based FR systems: one trained on the CASIA-WebFace database [19] and one trained on the MS-Celeb database [20] by de Freitas Pereira *et al.* [21].
- Open-source version of FaceNet [22] implemented and trained on the MS-Celeb database [20] by Sandberg [28].

TABLE I
DETAILS OF DATABASES USED IN OUR EXPERIMENTS

Database	Year	No. of images	Resolution
Flickr-Faces-HQ [17]	2019	70,000	1024 × 1024
CASIA-WebFace [19]	2014	494,414	256 × 256
MS-Celeb [20]	2016	10,490,534	Up to 300 × 300
Multi-PIE [37]	2009	755,370	3072 × 2048
LFW [39]	2007	13,233	Various
MOBIO [40]	2012	30,326	Various
IJB-A [41]	2015	5,712	Various

- DR-GAN [30] trained on a combination of the Multi-PIE database [37] and the CASIA-WebFace database [19].
- ArcFace [8] trained on the MS-Celeb database [20].

We used the two Inception-ResNet-v2 based FR systems and DR-GAN for generating master faces and all of the FR systems for evaluating master face attacks.[2] All FR systems were pretrained and obtained from the Bob toolbox [38].

2) Databases: Seven databases were used for four different purposes:

- *Training StyleGAN:* Flickr-Faces-HQ (FFHQ) database [17].
- *Training FR systems:* CASIA-WebFace [19], MS-Celeb [20], and Multi-PIE [37].
- *Running LVE algorithm:* Training set of LFW - Fold 1 database [39] aligned by funneling [42] and both male and female components of training set of MOBIO database [40].
- *Evaluating master faces:* Corresponding development (dev) and evaluation (eval) sets of LFW database [39] and MOBIO database [40] plus dev set of IARPA Janus Benchmark A (IJB-A) database [41].[3] The dev sets were used for threshold selection for the FR systems (which was based on the calculated equal error rates).

Details of the databases used are shown in Table I. There are no overlapping subjects between the databases used for training StyleGAN, training the FR systems, and running the LVE algorithm. This demonstrates that the LVE algorithm can work well even when its components use mutually exclusive databases.

We used the InsightFace library[4] to estimate the age and gender distributions of the databases used for training StyleGAN and the FR systems and for generating master faces. For the MOBIO database, we used its annotated gender information. We ignored the Multi-PIE database since it contributes only as an additional part of the database for training the DR-GAN FR system. The estimated distributions are shown in Fig. 7 and Fig. 8 respectively. The ages are dominantly 21 to 40, especially in the CASIA-WebFace, MS-Celeb, and MOBIO databases. The LFW - Fold 1 database is more balanced with a larger proportion of 41 to 60 ages. There are tiny numbers of child faces in all databases except for the MOBIO one, which has none. For gender, there are more male than female faces in all databases. The LFW - Fold 1 and MOBIO databases are the most unbalanced, with less than 25% female faces. This may cause bias in the FR systems as well as affect the properties of the generated master faces, as explained in the following section.

[2]Benchmarks for some of the systems can be found at https://www.idiap.ch/software/bob/docs/bob/bob.bio.face_ongoing/v1.0.4/leaderboard.html

[3]There is no eval set for the IJB-A database.

[4]https://github.com/deepinsight/insightface

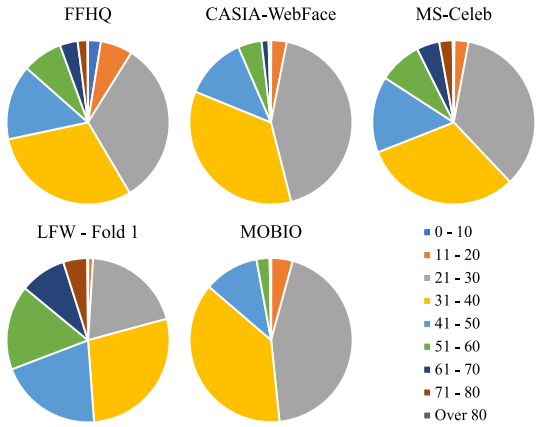

Fig. 7. Estimated age distribution of five databases used for training StyleGAN, FR systems, and generation of master faces. Best viewed in color.

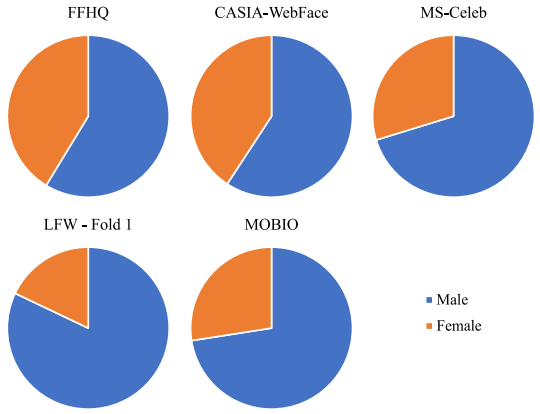

Fig. 8. Estimated gender distribution of five databases used for training StyleGAN, FR systems, and generation of master faces. Best viewed in color.

B. Latent Variable Evolution Configurations

Since there are many FR systems and databases, evaluating all possible combinations is impossible with the available computation and time resources. We thus selected a subset with the aim of covering a range as broad as possible. We defined eight settings (Table II) for the LVE algorithm using three FR systems (two versions of Inception-ResNet-v2, one trained on the CASIA-WebFace database and one trained on the MS-Celeb one, and DR-GAN) and two databases (LFW - Fold 1 and MOBIO). There are five settings in which one FR system and one database are used (*single 1* to *single 5*) and three settings in which more than one FR system and/or database is used (*combination 1*, *combination 2*, and *combination 3*).

Each combination setting combined two single settings and was selected on the basis of its reasonable coverage of cases. The main differences among the three combination settings are highlighted in Table III. In the *combination 1* setting, only one database was used with the LVE algorithm, and the databases used for training the FR systems were similar. In the

TABLE II
SETTINGS FOR RUNNING LVE ALGORITHM. "SINGLE" MEANS USING ONLY ONE DATABASE AND ONE FR SYSTEM. "COMBINATION" MEANS USING MORE THAN ONE DATABASE AND/OR FR SYSTEM. FOR EACH FR SYSTEM, WE SHOW BOTH ITS NETWORK ARCHITECTURE (TOP ROW) AND ITS TRAINING DATABASE (BOTTOM ROW)

No.	Setting	Database 1	FR System 1 (FR Training DB)	Database 2	FR System 2 (FR Training DB)
1	*Single 1*	LFW - Fold 1	Inception-ResNet-v2 (CASIA-WebFace)		
2	*Single 2*	LFW - Fold 1	DR-GAN (CASIA-WebFace & Multi-PIE)		
3	*Single 3*	MOBIO	Inception-ResNet-v2 (MS-Celeb)		
4	*Single 4*	LFW - Fold 1	Inception-ResNet-v2 (MS-Celeb)		
5	*Single 5*	MOBIO	DR-GAN (CASIA-WebFace & Multi-PIE)		
6	*Combination 1* (No. 1 & 2)	LFW - Fold 1	Inception-ResNet-v2 (CASIA-WebFace)	LFW - Fold 1	DR-GAN (CASIA-WebFace & Multi-PIE)
7	*Combination 2* (No. 1 & 3)	LFW - Fold 1	Inception-ResNet-v2 (CASIA-WebFace)	MOBIO	Inception-ResNet-v2 (MS-Celeb)
8	*Combination 3* (No. 4 & 5)	LFW - Fold 1	Inception-ResNet-v2 (MS-Celeb)	MOBIO	DR-GAN (CASIA-WebFace & Multi-PIE)

TABLE III
COMPARISON OF THREE COMBINATION SETTINGS FOR LVE ALGORITHM. FOR FR SYSTEMS, WE COMPARED THEIR ARCHITECTURES AND TRAINING DATABASES

Setting	Database 1 vs. Database 2	FR System 1 vs. FR System 2	
		Architectures	Training DBs
Combination 1	Same	Different	Similar
Combination 2	Different	Same	Different
Combination 3	Different	Different	Different

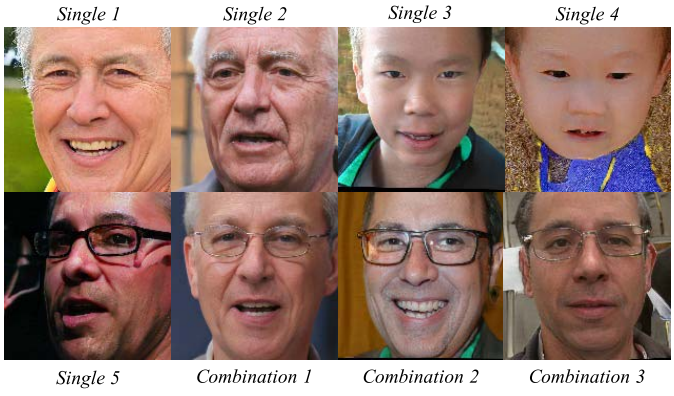

Fig. 9. Master faces generated using eight settings specified in Table II.

Fig. 10. Master face (top left) generated using *combination 1* setting and all matched faces from eval set of LFW - Fold 1 database [39] sorted from closest to farthest match. Inception-ResNet-v2 based FR system [21] was used in this case.

combination 2 setting, two databases were used with the LVE algorithm, and two FR systems with the same architecture but trained on different databases were used. In the *combination 3* setting, two databases and two FR systems without anything in common were used with the LVE algorithm. We ran 1000 iterations of the LVE algorithm for each of the eight settings.

The generated master faces corresponding to the eight settings are shown in Fig. 9. All of them are male faces. One-fourth are child faces, generated using only the Inception-ResNet-v2 based FR system trained on the MS-Celeb database. Half are elder faces generated using only the Inception-ResNet-v2 based FR system trained on the CASIA-WebFace database or only the DR-GAN FR system trained on the combination of the CASIA-WebFace and Multi-PIE databases, or a combination of these two FR systems. The rest (one-fourth) are middle-aged faces, generated using the combinations of the two FR systems in the previous two cases (one in each case).

C. Master Face Analysis

The master face generated using the *combination 1* setting and the faces it matched using the Inception-ResNet-v2 based FR system [21] on the eval set of the LFW - Fold 1 database [39] are shown in Fig. 10. The master face matched

those of persons of both genders, of multiple races (White, Black, and Asian), and of multiple ages (from children to elders). In many cases, the facial angles and lighting conditions differed from those of the master face. The subjects are both wearing and not wearing glasses (eyeglasses or sunglasses). A typical master face can match about 10 to 50 identities. Since the LFW database is unbalanced (as shown in Figs. 7 and 8), a large portion of the matched faces are male. Furthermore, since the master face falls in the elder cluster (discussed below), most of the matched faces are those of elders.

To better understand these results, we ran the uniform manifold approximation and projection (UMAP) dimension reduction algorithm on the embedding spaces of three FR systems and then applied a kernel density estimation method to the reduced spaces to form the density maps. We did that from both age and gender perspectives. We used two Inception-ResNet-v2 based FR systems (CASIA-WebFace version and MS-Celeb version) and the ArcFace FR system to perform the embedding space density estimation. Among them, the two Inception-ResNet-v2 FR systems were used on both the attacker side and the defender side while the ArcFace FR system was used only on the defender side. The estimated densities are shown in Fig. 11. We also included the positions of the intermediate master faces' and the optimized master faces' embeddings in the plots.

From the age perspective, young faces (less than 30 years old) are separated from the elder faces (more than 60 years old), while the remaining faces (30 to 60 years old) are scattered throughout both the young and elder faces. From the gender perspective, the male and female faces are somewhat separated. To maximize the false matches, the LVE algorithm placed the master face in a dense area near the border of a cluster, which increased the probability of matching diverse faces. Since there are more male than female faces in all databases, the probability of placement in a dense area in the male cluster was higher than that of placement in the female one. However, since they were only somewhat separated, the master faces could match both male and female faces (with more male face matches, as shown in Fig. 10).

For age, the selected dense area could be in a young cluster, a middle-aged cluster, or a elder cluster. Since the training data for the FR systems was unbalanced in terms of age with only a few samples for young and elder faces, these systems may not accurately recognize young and elder faces. The CASIA version of the Inception-ResNet-v2 based FR system may perform poorly on elder male faces, resulting in the generation of elder male master faces. Interestingly, the master face generated using the *combination 1* setting also lies at the centroid of the ArcFace FR system, which is used only on the defender side. For this case, dense areas also exist even if we use the angular margin loss in training.

On the other hand, the MS-Celeb version of the Inception-ResNet-v2 based FR system performed poorly on young male faces, resulting in the generation of boy master faces. For *combination 2* (not fully shown in Fig. 11 due to limited space), we observed that the 30- to 60-year-old faces were scattered in the embedding spaces of both of these FR systems; it seems that an "average" middle-aged face is the optimal solution according to the proposed LVE algorithm. To further verify the effects of the clusters on the properties of the master faces, we generated two master faces using only the female part of the MOBIO database and the Inception-ResNet-v2 FR system (MS-Celeb version) and the DR-GAN FR system. Both master faces are female, as shown in Fig 12.

D. False Matching Rate Analysis

Next, we evaluated the performances of attacks using master faces. The greater the number of enrolled subjects that match the generated master face, the higher the FMR. Hence, we compared the FMRs between two tests:

- *Normal test:* One side of the test pairs included either a genuine or zero-effort imposter face defined by the test protocols of the database used.
- *Master face test:* The master face was paired with the faces of all the enrolled subjects.

First, we show how the FMRs measured on the master face set changed during the LVE optimization. As shown in Fig. 13, the FMRs became higher in six of the eight settings. For the two remaining settings (*combination 2* and *3*), the FMR of one of their component FR systems also became higher while that of the other one remained almost zero. In these two cases, two different databases were used with the LVE algorithm, and the algorithm tried to maximize the similarities between the master face and all faces in database 1 as calculated by component FR system 1 as well as to maximize the similarities between the master face and all faces in database 2 as calculated by component FR system 2. This task is difficult, even if the two FR systems share the same architecture, as they do in the *combination 2* setting. Since the LFW and MOBIO databases have different distributions, finding a master face that matches the face of many subjects in both of them is challenging. The LVE algorithm focused on only one database (the LFW database) and ignored the other (the MOBIO database, which has higher variability in terms of pose and illumination conditions than the LFW database). Moreover, the Inception-ResNet-v2 based FR system trained on the MS-Celeb database was harder to fool when it was run with the LVE algorithm compared with its CASIA-WebFace version. In contrast, although two FR systems were used in the *combination 1* setting, they shared the same database, so the algorithm was able to fool both of them.

Two rules for designing settings for the LVE algorithm can be inferred from these results:

- Using more than one database for running the LVE algorithm is difficult as the algorithm may prioritize the database that is less challenging.
- Using more than two FR systems is OK. They can have the same or different architectures, trained on similar or different databases.

Table IV shows the FMRs for the normal tests and the corresponding master face tests using master faces generated using five *single* settings and three *combination* settings. Each cell has four numbers, the FMRs for the normal test (upper part) and the master face test (lower part), from the development set (left) and the evaluation set of the target database (right). Gray

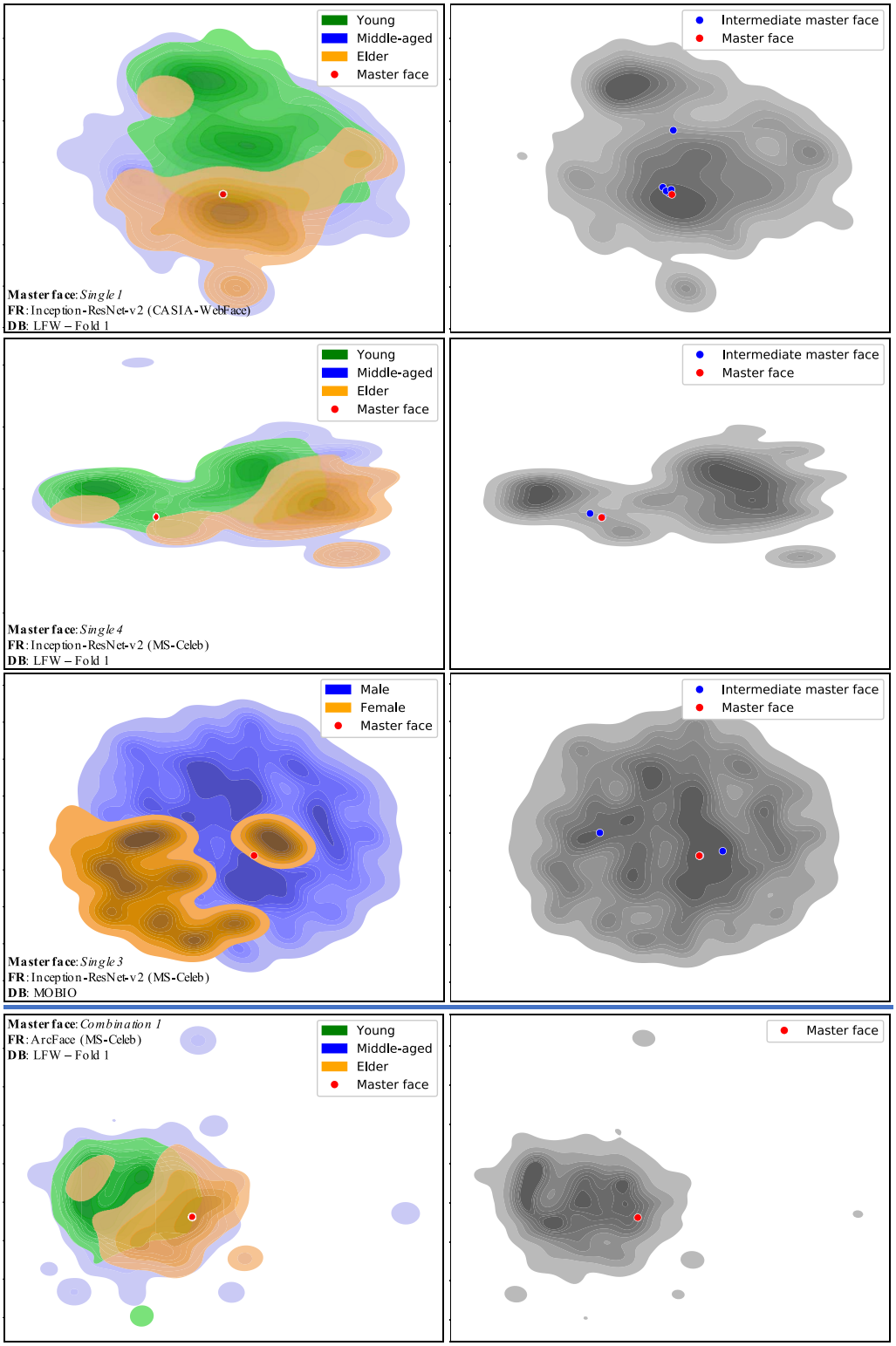

Fig. 11. Estimated densities of ages (rows 1, 2, and 4) and genders (row 3) of embedded faces extracted by Inception-ResNet-v2 based FR systems trained on CASIA-WebFace database (row 1) and MS-Celeb database (rows 2 and 3) and by ArcFace FR system (row 4). Plots on left show estimated densities per class while those on right show estimated densities of all embeddings. Two Inception-ResNet-v2 based FR systems were used on both attacker and defender sides while ArcFace system was used only on defender side. We also included the embeddings of five intermediate master faces generated during running of LVE algorithm (blue dots) and of optimized master face (red dot). These embeddings were extracted from the entire training set of the LFW - Fold 1 database (rows 1, 2, and 4) and of the MOBIO database (row 3). Corresponding LVE settings (see Table II for more detail) used to generate master faces are shown in figures on left, along with information about target FR system (denoted as FR) and database (denoted as DB). Best viewed in color.

cells indicate that the surrogate database(s) used by attackers when running the LVE algorithm and the target database(s) were different while gray cells indicate that they were the same. Numbers in bold indicate successful master face attacks. There are several observations regarding the FMRs of the attacks using the master faces generated using the *single* and

TABLE IV
FMRS OF NORMAL TESTS AND CORRESPONDING MASTER FACE TESTS USING MASTER FACES GENERATED USING FIVE *Single* SETTINGS AND THREE *Combination* SETTINGS. FOR EACH FR SYSTEM, WE SHOW BOTH ITS NETWORK ARCHITECTURE (TOP ROW) AND ITS TRAINING DATABASE (BOTTOM ROW). WITHIN EACH CELL, NUMBER AT UPPER LEFT IS FMR FOR NORMAL TEST FROM DEVELOPMENT SET OF TARGET DATABASE, THAT AT UPPER RIGHT IS FMR FOR NORMAL TEST FROM EVALUATION SET OF TARGET DATABASE, THAT AT LOWER LEFT IS FMR FOR MASTER FACE TEST FROM DEVELOPMENT SET OF TARGET DATABASE, AND THAT AT LOWER RIGHT IS FMR FOR MASTER FACE TEST FROM EVALUATION SET OF TARGET DATABASE. GRAY CELLS INDICATE THAT SURROGATE DATABASE(S) USED BY ATTACKERS WHEN RUNNING LVE ALGORITHM AND TARGET DATABASE(S) WERE DIFFERENT WHILE WHITE CELLS INDICATE THAT THEY WERE THE SAME. NUMBERS IN *italics* INDICATE THAT SURROGATE FR SYSTEM(S) AND TARGET FR SYSTEM(S) WERE IDENTICAL IN BOTH ARCHITECTURE(S) AND TRAINING DATABASE(S). NUMBERS IN BOLD INDICATE SUCCESSFUL MASTER FACE ATTACKS.

Target DB	Target FR System	*Single 1*		*Single 2*		*Single 3*		*Single 4*		*Single 5*		*Comb. 1*		*Comb. 2*		*Comb. 3*	
LFW - Fold 1	Inception-ResNet-v2	*2.3*	*3.3*	2.3	3.3	2.3	3.3	2.3	3.3	2.3	3.3	*2.3*	*3.3*	*2.3*	*3.3*	2.3	3.3
	(CASIA-WebFace)	***29.9***	***34.7***	**15.4**	**19.7**	2.4	2.3	0.4	0.4	0.8	0.2	***26.6***	***29.7***	***23.4***	***27.3***	**5.1**	4.4
	Inception-ResNet-v2	0.5	0.3	0.5	0.3	*0.5*	*0.3*	*0.5*	*0.3*	0.5	0.3	0.5	0.3	*0.5*	*0.3*	*0.5*	*0.3*
	(MS-Celeb)	0.1	0.8	1.0	1.1	***7.3***	***5.7***	***10.0***	***8.1***	1.1	0.6	1.2	1.7	*2.0*	*2.3*	*2.3*	*0.8*
	FaceNet (Inception-v1)	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3	0.7	0.3
	(MS-Celeb)	0.0	1.1	0.3	1.1	0.4	0.8	0.5	1.5	0.4	0.2	1.3	0.6	0.8	1.3	1.2	0.6
	DR-GAN	3.3	3.7	*3.3*	*3.7*	3.3	3.7	3.3	3.7	*3.3*	*3.7*	*3.3*	*3.7*	3.3	3.7	*3.3*	*3.7*
	(CASIA-WebFace)	**6.2**	**8.1**	***30.1***	***33.3***	1.1	0.8	2.0	0.8	*4.3*	*3.6*	***27.3***	***27.8***	3.5	2.8	***6.0***	***5.5***
	ArcFace	14.3	12.3	14.3	12.3	14.3	12.3	14.3	12.3	14.3	12.3	14.3	12.3	14.3	12.3	14.3	12.3
	(MS-Celeb)	11.6	13.6	**21.3**	**26.3**	2.4	2.5	4.9	2.5	9.7	7.6	**22.1**	**23.5**	14.3	15.3	13.6	14.6
MOBIO	Inception-ResNet-v2	*1.9*	*2.1*	1.9	2.1	1.9	2.1	1.9	2.1	1.9	2.1	*1.9*	*2.1*	*1.9*	*2.1*	1.9	2.1
	(CASIA-WebFace)	*2.4*	*0.0*	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*2.4*	*0.0*	*0.0*	*0.0*	**4.8**	**5.2**
	Inception-ResNet-v2	1.0	0.4	1.0	0.4	*1.0*	*0.4*	*1.0*	*0.4*	1.0	0.4	1.0	0.4	*1.0*	*0.4*	*1.0*	*0.4*
	(MS-Celeb)	0.0	0.0	0.0	0.0	*0.0*	*0.0*	*0.0*	*0.0*	0.0	0.0	0.0	0.0	*0.0*	*0.0*	*0.0*	*0.0*
	FaceNet (Inception-v1)	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5
	(MS-Celeb)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.0	0.0	0.0	0.0
	DR-GAN	2.3	1.3	*2.3*	*1.3*	2.3	1.3	2.3	1.3	*2.3*	*1.3*	*2.3*	*1.3*	2.3	1.3	*2.3*	*1.3*
	(CASIA-WebFace)	0.0	0.0	*0.0*	*0.0*	0.0	0.0	0.0	0.0	*2.4*	***12.1***	*0.0*	*0.0*	0.0	0.0	***4.8***	***6.9***
	ArcFace	8.2	7.8	8.2	7.8	8.2	7.8	8.2	7.8	8.2	7.8	8.2	7.8	8.2	7.8	8.2	7.8
	(MS-Celeb)	0.0	0.0	2.4	1.7	0.0	0.0	0.0	0.0	0.0	0.0	4.8	1.7	0.0	0.0	4.8	3.4
IJB-A	Inception-ResNet-v2	*10.1*		10.1		10.1		10.1		10.1		*10.1*		*10.1*		10.1	
	(CASIA-WebFace)	***37.5***		**15.2**		**13.4**		9.8		2.7		***20.5***		***25.0***		2.7	
	Inception-ResNet-v2	3.1		3.1		*3.1*		*3.1*		3.1		3.1		*3.1*		*3.1*	
	(MS-Celeb)	0.0		1.8		***20.5***		***22.3***		0.0		1.8		*0.9*		*0.0*	
	FaceNet (Inception-v1)	5.7		5.7		5.7		5.7		5.7		5.7		5.7		5.7	
	(MS-Celeb)	4.5		2.7		**9.8**		**8.0**		1.8		6.2		4.5		4.5	
	DR-GAN	10.8		*10.8*		10.8		10.8		*10.8*		*10.8*		10.8		*10.8*	
	(CASIA-WebFace)	4.5		***16.1***		7.1		**14.3**		*5.4*		***17.9***		5.4		*1.8*	
	ArcFace	10.7		10.7		10.7		10.7		10.7		10.7		10.7		10.7	
	(MS-Celeb)	0.0		3.6		0.9		0.9		2.7		0.9		0.9		0.9	

Fig. 12. Two female master faces generated using only female part of MOBIO database and Inception-ResNet-v2 based FR system (MS-Celeb version) and DR-GAN FR system, respectively.

TABLE V
SUMMARY OF SUCCESSFUL ATTACK RATIOS USING **FIVE** *Single* SETTINGS AND **THREE** COMBINATION SETTINGS. NUMERATORS ARE NUMBER OF SUCCESSFUL ATTACKS; DENOMINATORS ARE TOTAL NUMBER OF ATTACK CASES. NOTE THAT FOR *Combination 3*, SOME ATTACKS FALL INTO TWO SETTINGS: "SAME ARCH. - DIFFERENT DB" AND "DIFFERENT ARCH. - SAME DB." NUMBERS FOR OVERLAPPED CASES ARE SHOWN INSIDE PARENTHESES

Target FR (Architecture - Training DB)	*Single* Settings		*Combination* Settings	
	Known target DB	Unknown target DB	Known target DB	Unknown target DB
Same Arch. - Same DB	**7/10**	**6/15**	**10/20**	**3/10**
Same Arch. - Different DB	0/6	**1/9**	**3/6 (3/4)**	0/4 (0/1)
Different Arch. - Same DB	**4/14**	**3/21**	**3/24 (3/4)**	0/6 (0/1)
Different Arch. - Different DB	**2/20**	**1/30**	**2/4**	0/6
Overall success ratio	0.26	0.15	0.30	0.12
	0.19		0.24	
	0.21			

combination settings shown in Table IV in connection with the FMR curves shown in Fig. 13.

- All FR systems are vulnerable to master face attacks. Some systems are easier to fool than others.
- With the *combination 1* setting, the master face had the attack abilities of the master faces generated using the corresponding single settings (*single 1* and *single 2*). In this case, there was no conflict.
- With the *combination 2* and *combination 3* settings, in which conflict occurred, their master faces were lacking some attack abilities of the master faces generated using the corresponding single settings. This is clearly seen for the *combination 2* setting, for which six attacks that were successful in the single settings failed.
- With the *combination 3* setting, for which the two component databases and FR systems differed, five attacks that had been successful were no long successful, and there were six newly successful ones. Moreover, the FMRs of the successful attacks were not as high as the those of the single setting. Although conflict still occurred in this case, it was less severe than in the *combination 2* setting.

The above observations provide valuable clues for effectively designing the LVE algorithm. Using only one database is a safe way to avoid conflicts when running the LVE algorithm. Although there negative side effects due to conflicts, using both different databases and different FR systems may result in unpredictable successful attacks when the single setting fails.

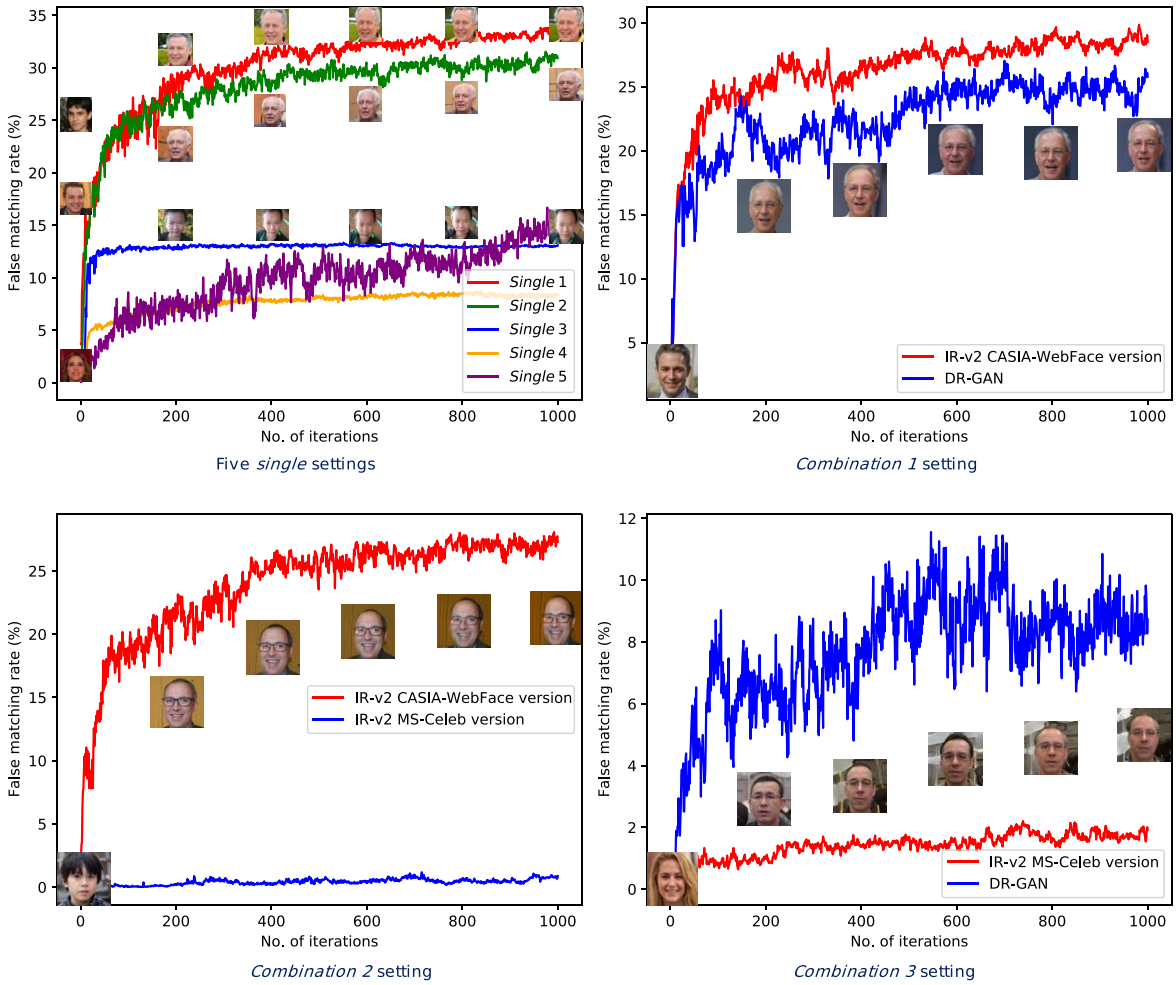

Fig. 13. FMRs of each FR system when running LVE algorithm using five *single* settings and three *combination* settings. Two Inception-ResNet-v2 (IR-v2) FR systems were used, one trained on CASIA-WebFace database and one trained on MS-Celeb database. We included intermediate master faces generated using three *single* settings (1, 2, and 3) and three combination settings. Best viewed in color.

Table V summarizes the number of successful attacks using both *single* and *combination* settings. An attack is successful if the master face's FMR is higher than the normal test set's FMR. Recall that there were **five** *single* settings and **three** *combination* settings. Moreover, the *combination* settings used more than one database and/or one FR system, and there were only three databases and five FR systems used for evaluation. As a result, the total number of black/gray-box attacks (attacks on different architecture, different database) with these settings was less than that of attacks with the *single* settings.

The overall success ratio of master face attacks was 21%. White-box attacks had the highest success ratios, followed sequentially by gray-box and black-box attacks. The success rate for the *combination* settings (24%, overlapped cases removed) is higher than that for the *single* settings (19%). This means that, although the generation process is more difficult for the *combination* settings, when the attacks are successful, the master faces have stronger attack ability. Regarding black-box attacks (both target database and FR system are unknown), since we had only a limited number of scenarios (6 in total for *combination* settings compared with 30 for *single* settings), it is hard to conclude whether master faces generated using *combination* settings can successfully perform black-box attacks. The main point of using a *combination* setting is to increase the chance of an attack being a gray-box or white-box (if lucky) attack by using multiple databases and FR systems for guessing and approximating the target system.

In reality, attackers can mix several databases to create a single large database with increased generalizability. There are not many public FR system architectures; therefore, attackers can prepare in advance several master faces for each one using a mixed database.

V. Presentation Attacks

Finally, we evaluated the risk and threats of presentation attacks using master faces on FR systems. For master face candidates, we chose one generated using the *single 2* setting and another generated using the *combination 1* setting. For digital attack candidates, we chose two attack scenarios in the IJB-A database [41] in which the two master faces were falsely accepted by the Inception-ResNet-v2 based FR system [21] (CASIA-WebFace version) and the DR-GAN FR system [30].

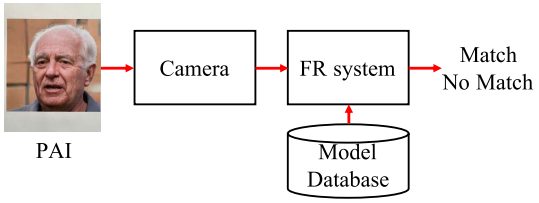

Fig. 14. Overview of presentation attack on FR system.

TABLE VI
FMRs OF MASTER FACE PAI ATTACKS ON DEV SET OF IJB-A DATABASE [41] USING TWO SETTINGS: *Single 2* AND *Combination 1*. FIRST LINE IN EACH ROW SHOWS RESULT FOR INCEPTION-RESNET-V2 BASED FR SYSTEM [21] TRAINED ON CASIA-WEBFACE DATABASE, AND SECOND LINE SHOWS RESULT FOR DR-GAN FR SYSTEM [30]. NUMBERS IN **BOLD FONT** INDICATE SUCCESSFUL ATTACKS ALTHOUGH THERE WAS DEGRADATION IN THE FMR IN SOME CASES

Camera	Plain Paper	Photo Paper	MacBook Screen	Digital Master Face	Normal Dev Set
	Setting: *Single 2*				
iPhone XR	**17.9 (+2.7)**	**15.2 (0.0)**	**13.4 (−1.8)**	15.2	10.1
	18.8 (+2.7)	11.6 (−4.5)	**19.6 (+3.5)**	16.1	10.8
Canon 60D	**17.9 (+2.7)**	**18.8 (+3.6)**	**19.6 (+4.4)**	15.2	10.1
	20.5 (+4.4)	11.6 (−4.5)	**15.2 (−0.9)**	16.1	10.8
	Setting: *Combination 1*				
iPhone XR	**18.8 (−1.7)**	**19.6 (−0.9)**	**15.2 (−5.3)**	20.5	10.1
	13.4 (−4.5)	11.6 (−6.3)	8.9 (−9.0)	17.9	10.8
Canon 60D	**17.9 (−2.6)**	**19.6 (−0.9)**	**20.5 (0.0)**	20.5	10.1
	13.4 (−4.5)	8.0 (−9.9)	**18.8 (+0.9)**	17.9	10.8

We compared the FMRs of these two digital attacks with those of the corresponding presentation attacks.

A. Experiment Design

To simulate simple presentation attacks like the one shown in Fig. 14, we needed to prepare PAIs and cameras. For the PAIs of each of the two selected master faces, we used three kinds of materials:

- Color photos printed on plain A4 paper.
- Color photos printed on 127 mm × 178 mm photo paper.
- Color photos displayed on the screen of an Apple 13-inch MacBook Pro 2017.

For the cameras, we used two types:

- the rear camera in an iPhone XR.
- a Canon EOS 60D DSLR camera with a Canon EF 40mm F2.8 STM lens.

For simplicity, we used these cameras to take photos of the PAIs under normal room conditions. We adjusted the position of the cameras such that they were relatively perpendicular to the surface of the PAIs so they could capture the displayed PAIs as much as possible without loosing any contents. This condition is close to that of real-world presentation attacks. Three example PAIs are shown in Fig. 1.

B. Results

The FMRs of the attacks using PAI master faces are shown in Table VI along with those of attacks using digital master faces and those of the normal dev set of the IJB-A database. The attacks were successful in 19 of the 24 cases, demonstrating that PAI master faces can be effective in real-world attacks. In eight cases, the FMRs were higher than those of attacks using digital master faces. This is attributed to the distribution of PAI master faces being closer to the distribution of faces in the facial databases (which contain faces also captured with a camera) thanks to the camera processing. The lower rate in the other cases is attributed to artifacts from the PAI materials playing a bigger role than the effect of the camera processing. All of the PAI attacks using plain paper were successful while seven of the eight PAI attacks using a computer screen were successful. The attacks using photo paper, which easily reflects light, had the worst performance. Those using photos taken with the iPhone camera were more successful than those using ones taken with the Canon camera. This is attributed to the Canon camera being able to capture more detailed PAI artifacts.

VI. DEFENSE AGAINST MASTER FACE ATTACKS

What is the main problem of existing FR systems that causes the existence of master faces? We hypothesized that it comes from the distributions of the embedding spaces where the extracted features are not well distributed. This results in the formation of clusters, not only multi-identity clusters but also age and gender ones. There are two possible origins of this problem: (1) the training data and (2) the objective function design. Regarding the training data, as shown in Figs. 7 and 8, the training data was unbalanced in terms of age and gender. This could affect the distribution of the embeddings for which the FR systems discriminate faces in the majority group better than in the minority one. For example, the 30-60 year-old face embeddings were scattered more uniformly than the others, as shown in Fig. 11.Simply enlarging the database has a certain effect on the robustness of the FR systems (the MS-Celeb version of the Inception-ResNet-v2 based FR system had fewer successful master face attacks than the CASIA-WebFace version); however, they are still vulnerable. It is thus important to balance the training data.

Regarding the objective function design, the objective functions are mainly designed so that same-identity embeddings stay close together while different-identity ones stay far apart. The introduction of the angular margin loss [8] improves this ability while the uniform loss [31] forces the embeddings to be uniformly distributed. Although these improvements reduce the risk of master face attacks, they mainly focus on identity. Since gender, age, and race are also important [43], the attack is successful in some cases. This suggests that the design of the objective functions used for training the FR systems needs further improvement.

Beside harnessing FR systems, using master face detectors could mitigate master face attacks. Since master faces are generated using a GAN, GAN image detectors [44]–[46] or deepfake detectors [47] could be used to detect them. Although looking realistic from the human perspective, computer-generated images have different properties than natural ones captured by cameras. Some GAN artifacts may exist in the generated images; therefore, most GAN image detectors focus on detecting their presence. We could also integrate a presentation attack detector [47] with an FR system to prevent master face attacks as well as other traditional presentation

attacks using images or videos of the victims. However, generalization of these detectors is still a huge challenge. The StyleGAN used in the LVE algorithm could be replaced with a more advanced facial generator to fool fake image detectors. Although some degree of generalizability has been achieved, performance is still not good enough for real-world applications. Therefore, further research on generalizability is needed.

VII. Conclusion

We have again demonstrated, especially in our presentation attack experiment, that master face attacks pose a severe security threat if the FR systems are not properly protected. Our intensive evaluation of the performance of the LVE algorithm using several settings, including both *single* and *combination* settings, has brought to light several properties of master faces as well as of the LVE algorithm. Some of the *combination* settings caused intra-component conflicts while others produced interesting positive results. Being aware of the existence of master faces and their properties is critical to improving the robustness of FR systems. Combining the use of an FR system with a well-designed objective function trained on a large balanced database with a fake image detector could mitigate master face attacks. Since digital attack detectors (GAN image detectors and deepfake detectors) and presentation attack detectors still have difficulty with generalization and since master face attacks continue to improve, these attacks cannot be taken lightly. Future work will focus on designing a better method to generate master faces and one to detect master face attacks.

Acknowledgment

The authors would like to thank Dr. Tiago de Freitas Pereira and Dr. Amir Mohammadi of the Biometrics Security and Privacy (BSP) group at the Idiap Research Institute for providing the pretrained face recognition systems and for their support with the Bob toolkit.

References

[1] "The Goode Intelligence Biometric Survey 2021." Goode Intelligence. Apr. 2021. [Online]. Available: https://www.goodeintelligence.com/report/the-goode-intelligence-biometric-survey-2021/

[2] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, "Recent advances in face presentation attack detection," in *Handbook of Biometric Anti-Spoofing*. Cham, Switzerland: Springer, 2019, pp. 207–228.

[3] P. Bontrager, W. Lin, J. Togelius, and S. Risi, "Deep interactive evolution," in *Proc. Int. Conf. Comput. Intell. Music Sound Art Des.*, 2018, pp. 267–282.

[4] H. H. Nguyen, J. Yamagishi, I. Echizen, and S. Marcel, "Generating master faces for use in performing wolf attacks on face recognition systems," in *Proc. IJCB*, 2020, pp. 1–10.

[5] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[6] P. Bontrager, A. Roy, J. Togelius, N. Memon, and A. Ross, "DeepMasterPrints: Generating MasterPrints for dictionary attacks via latent variable evolution," in *Proc. BTAS*, 2018, pp. 1–9.

[7] M. Une, A. Otsuka, and H. Imai, "Wolf attack probability: A new security measure in biometric authentication systems," in *Proc. ICB*, 2007, pp. 396–406.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.

[10] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[11] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, 2017, pp. 214–223.

[12] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5769–5779.

[13] H. Huang, Z. Li, R. He, Z. Sun, and T. Tan, "IntroVAE: Introspective variational autoencoders for photographic image synthesis," in *Proc. NIPS*, 2018, pp. 52–63.

[14] A. Razavi, A. van den Oord, and O. Vinyals, "Generating diverse high-fidelity images with VQ-VAE-2," in *Advances in Neural Information Processing Systems*, 2019, pp. 14866–14876. [Online]. Available: https://papers.nips.cc/paper/2019

[15] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR*, 2018.

[16] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. ICLR*, 2018.

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

[18] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020, pp. 8110–8119.

[19] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[20] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. ECCV*, 2016, pp. 87–102.

[21] T. de Freitas Pereira, A. Anjos, and S. Marcel, "Heterogeneous face recognition using domain specific units," *IEEE Trans. Inf. Forensics Security*, vol. 14, pp. 1803–1816, 2019.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, 2015, pp. 815–823.

[23] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, Apr. 2015.

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

[25] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 4278–4284.

[26] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. BMVC*, 2015, pp. 1–12.

[27] X. Wu, R. He, Z. Sun, and T. Tan, "A light CNN for deep face representation with noisy labels," *IEEE Trans. Inf. Forensics Security*, vol. 13, pp. 2884–2896, 2018.

[28] D. Sandberg. "FaceNet: Face Recognition Using Tensorflow." 2017. [Online]. Available: https://github.com/davidsandberg/facenet

[29] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014, pp. 1701–1708.

[30] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. CVPR*, 2017, pp. 1415–1424.

[31] Y. Duan, J. Lu, and J. Zhou, "UniformFace: Learning deep equidistributed representation for face recognition," in *Proc. CVPR*, 2019, pp. 3415–3424.

[32] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, "Introduction to face presentation attack detection," in *Handbook of Biometric Anti-Spoofing*. Cham, Switzerland: Springer, 2019, pp. 187–206.

[33] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Syst. J.*, vol. 40, no. 3, pp. 614–634, 2001.

[34] M. Inuma, A. Otsuka, and H. Imai, "Theoretical framework for constructing matching algorithms in biometric authentication systems," in *Proc. ICB*, 2009, pp. 806–815.

[35] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evol. Comput.*, vol. 9, no. 2, pp. 159–195, 2001.

[36] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," *J. Open Sour. Softw.*, vol. 3, no. 29, p. 861, 2018.
[37] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
[38] A. Anjos, M. Günther, T. de Freitas Pereira, P. Korshunov, A. Mohammadi, and S. Marcel, "Continuously reproducing toolchains in pattern recognition and machine learning experiments," in *Proc. ICML*, Aug. 2017.
[39] G. B. Huang and E. Learned-Miller, "Labeled faces in the wild: Updates and new reporting procedures," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Rep. UM-CS-2014-003, May 2014.
[40] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. ICMEW*, 2012, pp. 635–640.
[41] B. F. Klare *et al.*, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A," in *Proc. CVPR*, 2015, pp. 1931–1939.
[42] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. ICCV*, 2007, pp. 1–8.
[43] P. Grother, M. Ngan, and K. Hanaoka, *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. Gaithersburg, MD, USA: Nat. Inst. Stand. Technol., 2019.
[44] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *Proc. WIFS*, 2019, pp. 1–6.
[45] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proc. ICCV*, 2019, pp. 7556–7566.
[46] N. Hulzebosch, S. Ibrahimi, and M. Worring, "Detecting CNN-generated facial images in real-world scenarios," in *Proc. CVPR Workshops*, 2020, pp. 642–643.
[47] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020.

Huy H. Nguyen (Member, IEEE) received the Ph.D. degree from the Graduate University for Advanced Studies, SOKENDAI, Japan, in 2022. He is currently a Postdoctoral Researcher with the National Institute of Informatics, Tokyo, Japan. His research interests include security and privacy in biometrics and machine learning.

Sébastien Marcel (Senior Member, IEEE) is the Head of the Biometrics Security and Privacy Group, Idiap Research Institute (CH) and conducts research on face recognition, speaker recognition, vein recognition, and presentation attack detection. He is a Professor with the University of Lausanne and a Lecturer with the École Polytechnique Fédérale de Lausanne. He is also the Director of the Swiss Center for Biometrics Research and Testing, conducting FIDO certifications and research. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR, AND IDENTITY SCIENCE. He coordinated several European research projects, including MOBIO and TABULA RASA or BEAT and is involved in international projects (DARPA and IARPA).

Junichi Yamagishi (Senior Member, IEEE) received the Ph.D. degree from the Tokyo Institute of Technology (Tokyo Tech), Tokyo, Japan, in 2006.

From 2007 to 2013 he was a Research Fellow with the Centre for Speech Technology Research, University of Edinburgh, U.K. He became an Associate Professor with the National Institute of Informatics, Japan, in 2013, where he is currently a Professor. His research interests include speech processing, machine learning, signal processing, biometrics, digital media cloning, and media forensics. He served as a Co-Organizer for the bi-annual ASVspoof Challenge and the bi-annual Voice Conversion Challenge. He also served as a member for the IEEE Speech and Language Technical Committee from 2013 to 2019, as an Associate Editor for the IEEE/ACM TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING from 2014 to 2017, and as the Chairperson for ISCA SynSIG from 2017 to 2021. He is currently a Principal Investigator of the JST-CREST and ANR supported VoicePersona Project and a Senior Area Editor of the IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING.

Isao Echizen (Member, IEEE) received the B.S., M.S., and D.E. degrees from the Tokyo Institute of Technology, Japan, in 1995, 1997, and 2003, respectively. In 1997, he joined Hitachi Ltd., where he was a Research Engineer with the Systems Development Laboratory until 2007. He is currently the Director and a Professor with the Information and Society Research Division, National Institute of Informatics, and a Professor with the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan. He was a Visiting Professor with Tsuda University, Japan; University of Freiburg, Germany; and University of Halle–Wittenberg, Germany. He is currently engaged in research on multimedia security and multimedia forensics and serves as the Research Director for the CREST FakeMedia Project, Japan Science and Technology Agency. He was a member of the Information Forensics and Security Technical Committee and the IEEE Signal Processing Society. He is the Japanese representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems), a Member-at-Large of the APSIPA Board of Governors, and an Editorial Board Member of the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING and the *EURASIP Journal on Image and Video Processing*.